The AI Layoff Trap

Brett Hemenway Falk, Gerry Tsoukalas*

March 2, 2026

Abstract

If AI displaces human workers faster than the economy can reabsorb them, it risks eroding the very consumer demand firms depend on. We show that knowing this is not enough for firms to stop it. In a competitive task-based model, demand externalities trap rational firms in an automation arms race, displacing workers well beyond what is collectively optimal. The resulting loss harms both workers and firm owners. More competition and "better" AI amplify the excess; wage adjustments and free entry cannot eliminate it. Neither can capital income taxes, worker equity participation, universal basic income, upskilling, or Coasian bargaining. Only a Pigouvian automation tax can. The results suggest that policy should address not only the aftermath of AI labor displacement but also the competitive incentives that drive it.

Keywords: artificial intelligence, automation, labor displacement, Pigouvian tax.

1 Introduction

The fear that technology will displace workers is at least as old as the Industrial Revolution (Ricardo, 1821; Keynes, 1930; Leontief, 1982). Historically, displacement has largely been self-correcting: automation of existing tasks has been offset by the creation of new tasks and occupations. What Acemoglu and Restrepo (2018, 2019) call the reinstatement effect has tended to stabilize the labor market. Whether this balance will hold in the age of AI is an open question: Autor et al. (2024) find that displacement has intensified over the past four decades while the creation of new work has not always kept pace, and early signs suggest the current wave is disproportionately affecting entry-level workers (Brynjolfsson et al., 2025a).

Even if reinstatement eventually occurs, a problem arises along the way: displaced workers are also consumers, and when their lost income is not replaced, each round of layoffs erodes the purchasing power all firms depend on. At the limit, this becomes self-destructive: firms automate

---

*Hemenway Falk: University of Pennsylvania (fbrett@cis.upenn.edu), Tsoukalas: Boston University (gerryt@bu.edu)

their way to boundless productivity and zero demand. Public discourse increasingly treats this dynamic as an inevitable process with no natural brake (Shah, 2026). But rational, forward-looking firms *should* be the brake; if the cliff ahead is visible to all, why would they race toward it?

Yet the evidence suggests firms are heading in precisely that direction. In February 2026, Block cut nearly half its 10,000-person workforce, with CEO Jack Dorsey stating that AI had made many of those roles unnecessary and that "within the next year, the majority of companies will reach the same conclusion" (CNBC, 2026b). Over 100,000 tech workers were laid off in 2025 alone, with AI cited as a primary driver in more than half the cases, concentrated in customer support, operations, and middle management (CNBC, 2025b).[1] The exposure extends beyond tech: Eloundou et al. (2024) estimate that roughly 80% of U.S. workers hold jobs with tasks susceptible to automation by large language models. None of this is hidden. Against this backdrop, we ask under what conditions rationality and perfect foresight are enough to prevent competitive over-automation, what determines the size of the distortion when they are not enough, and which proposed policy responses correct it.

To answer these questions, we develop a task-based automation model inspired by Acemoglu and Restrepo (2018), but refocused from the labor market to the product market: when automation displaces workers, their forgone spending reduces every firm's revenue. Each of several symmetric firms chooses what fraction of its workforce to replace with AI. Automated tasks are performed at lower cost, but integration frictions make each successive task harder to automate. On the demand side, workers spend a fraction of their income on the sector's output; firm owners spend less, normalized to zero in the baseline. Some displaced wage income is recovered through reemployment or transfers, but the remainder is lost to the sector. The model is deliberately stripped down to make this channel transparent, and the demand cliff ahead visible to all firms. The baseline holds wages fixed and shuts down capital-income recycling; extensions relax both and several other baseline assumptions. Despite its parsimony, the framework accommodates a range of policy instruments and robustness checks.

We show that competition creates a demand externality that traps firms: an automating firm captures the full cost saving but, under competitive pricing, bears only a fraction of the resulting aggregate demand destruction; the rest falls on rivals. Each firm's profit-maximizing automation rate is a strictly dominant strategy that exceeds the cooperatively efficient level, so foresight alone cannot prevent the race toward the cliff. The distortion deepens with competition: a monopolist

---

[1] Individual cases illustrate the scale: Salesforce replaced 4,000 customer-support agents with agentic AI (CNBC, 2025c), and Cognition's Devin, deployed at Goldman Sachs and Infosys, enables one senior engineer to do the work of a five-person team (CNBC, 2025a; Infosys, 2026).

fully internalizes the externality, while fragmented markets exhibit the widest gap. In the frictionless limit, where every task is equally easy to automate, the game sharpens into a Prisoner's Dilemma in which every firm displaces its entire human workforce with AI, even though collective restraint would raise all profits. The resulting surplus loss is not a transfer from workers to firm owners; it is a deadweight loss that harms both.

Since the loss falls on both sides, a natural question is whether policy can correct it. We evaluate six instruments against the externality margin. Upskilling and worker equity participation narrow the wedge but cannot eliminate it. Nor can Coasian bargaining: because automation is a dominant strategy, no voluntary agreement among firms is self-enforcing. Capital income taxes do not alter the equilibrium automation rate, operating on profit levels rather than the per-task margin where the externality resides. Neither does universal basic income: it raises the floor on living standards but leaves the automation incentive unchanged. Only a Pigouvian automation tax, set equal to the uninternalized demand loss per task, implements the cooperative optimum; its revenue can fund retraining that raises income replacement, shrinking the externality over time and making the tax potentially self-limiting.

The core result is also robust to several generalizations. Higher AI productivity widens the wedge rather than resolving it: each firm perceives a market-share gain from automating beyond rivals, but at the symmetric equilibrium these gains cancel, leaving only the additional distortion. This Red Queen effect means that "better" AI, far from mitigating the externality, amplifies it. Endogenous wage adjustment, a key self-correcting channel in the framework of Acemoglu and Restrepo (2018), raises the threshold at which the externality activates but cannot close the wedge once it does: wage flexibility changes *when* the problem bites, not *whether* it exists. Free entry, capital-income recycling, and richer product-market structures likewise fail to eliminate the distortion.

Our work contributes to several literatures. We build on the task-based approach to automation (Zeira, 1998; Autor et al., 2003; Acemoglu and Restrepo, 2018, 2019), which emphasizes offsetting forces that restore labor demand after displacement, notably new task creation and a self-correcting wage channel. Acemoglu (2025) evaluates the aggregate productivity effects of AI within this framework. These contributions focus on whether and how the labor market rebalances; we ask what happens on the product-market side when rebalancing is slow or incomplete.

A growing literature argues that automation may be excessive. The closest to our setting is Beraja and Zorzi (2025), who show that automation is inefficient when displaced workers face borrowing constraints during reallocation. Their mechanism operates through the labor market: firms ignore the welfare cost imposed on credit-constrained workers. Ours operates through the product market: firms ignore the demand they destroy for rival firms. Their inefficiency arises even for a single firm in isolation; ours requires competition and vanishes under monopoly. And whereas their planner corrects automation to protect worker welfare, ours would reduce automation even with zero weight on workers, because over-automation harms firm profits themselves.

Other channels for excessive automation share the feature that they would distort a single firm's decision even in isolation: the technology ecosystem may be biased toward "so-so" automation that displaces workers without large productivity gains (Acemoglu and Restrepo, 2020), automation may disproportionately target high-rent tasks, dissipating worker surplus rather than raising output (Acemoglu and Restrepo, forthcoming), and corrective taxation has been justified by transitional frictions (Guerreiro et al., 2022) and distributional concerns (Costinot and Werning, 2023). Our externality, by contrast, arises only under competition and persists even when automation is highly productive, credit markets are complete, and the planner places no weight on distribution.

The demand externality we study belongs to the family of aggregate demand spillovers introduced by Rosenstein-Rodan (1943) and formalized by Murphy et al. (1989). In their "big push" models, demand complementarities across sectors can prevent individually unprofitable investments from being made even though simultaneous adoption would be collectively profitable. Our mechanism is the mirror image: *individually profitable* automation is collectively destructive because each firm's cost saving erodes the revenue base all firms share. Cooper and John (1988) provide the canonical framework for coordination failures driven by aggregate demand externalities; our game shares this structure but yields a unique dominant-strategy equilibrium, making the problem a true externality rather than a coordination failure that communication could resolve. Related work on automation and demand, including Benzell et al. (2015) on robot adoption in an overlapping-generations setting and Korinek and Stiglitz (2019) on AI-driven income redistribution, does not model the strategic interaction among firms that produces the externality we identify.

The information systems literature has established that AI systems deliver substantial productivity gains (Brynjolfsson et al., 2025b; Brynjolfsson and McAfee, 2014) and are increasingly deployed in strategic roles such as pricing, where algorithms can spontaneously learn to collude (Banchio and Mantegazza, 2022; Keppo et al., 2026). On the adoption side, Li et al. (2025) show that firms under labor-issue scrutiny invest specifically in AI automation rather than other forms of IT, and Bastani and Cachon (2025) show that as AI reliability improves, incentivizing effective human oversight becomes prohibitively expensive, weakening a key check on automation. What this literature has not modeled is how these individually documented phenomena interact across firms: each adoption decision is rational in isolation, but collectively they erode the consumer demand all firms depend on. We provide that model, connecting the micro-level evidence the IS literature has documented to a macro-level market failure that no individual firm can prevent.

The remainder of the paper is organized as follows. Section 2 presents the model. Section 3 derives the equilibrium and the over-automation wedge. Section 4 evaluates policy instruments. Section 5 extends the model to AI productivity gains, endogenous entry, endogenous wages, capital-income recycling, and richer product-market interaction. Section 6 discusses implications and limitations.

2 Model

The baseline isolates the demand consequences of automation in the simplest environment that supports the mechanism: symmetric firms, a single sector, and exogenous wages. We describe the supply side (cost structure and automation choice), then the demand side (how displacement feeds back into revenue), and finally the game firms play. Each assumption is relaxed in Section 5.

Consider a sector with $N \geq 2$ symmetric firms, indexed $i = 1, \ldots, N$. It will later become useful to think of each firm as having a single *owner*, for example the equity holder, who is entitled to the firm's operating profits.

In the spirit of the task-based framework of Acemoglu and Restrepo (2018), each firm is endowed with $L > 0$ task-positions. Initially all tasks are performed by human *workers*; a new technology shock arrives, for example agentic AI, and each firm must decide how much of its workforce to replace. In particular, firm i chooses an automation rate $\alpha_i \in [0, 1]$: tasks $z \in [0, \alpha_i]$ are performed by AI at cost c per task, and tasks $z \in (\alpha_i, 1]$ remain with human workers at wage w per task, with $0 \leq c \leq w$. Since each automated task displaces one worker, α_i is simultaneously the automation rate and the fraction of the workforce laid off; we use the two descriptions interchangeably. Wages are exogenous in the baseline; Section 5.3 endogenizes wages.

In the perfect-substitutes limit of the CES task aggregator in Acemoglu and Restrepo (2018), each task produces one unit of output regardless of mode, so firm output is $Y_i = L$; Section 5.1 relaxes this to allow AI to not only reduce costs, but also increase firm output. This normalization shuts down productivity and quality margins so that the baseline captures only the spending consequences of labor displacement.

We follow the literature in assuming tasks are ordered by comparative advantage, making the marginal task progressively harder to integrate; we capture this via a convex integration cost $\frac{k}{2} L \alpha_i^2$ with $k \geq 0$, using the standard quadratic adjustment-cost specification (Lucas, 1967; Hamermesh and Pfann, 1996). Firm i's total production cost is therefore

$$C_i(\alpha_i) = L\big(\alpha_i c + (1-\alpha_i)w\big) + \tfrac{k}{2}\, L\, \alpha_i^2. \tag{1}$$

Defining the per-task cost saving from automation as $s := w - c$, the cost equation can be rewritten as $C_i = L(w - s\,\alpha_i) + \frac{k}{2}\, L\, \alpha_i^2$: each automated task saves s in labor costs but incurs the integration friction.

On the demand side, workers have a higher marginal propensity to consume (MPC) than owners (Kaldor, 1956; Mian et al., 2021); workers spend a fraction $\lambda \in (0, 1]$ of their income on the sector's good, generating the type of cross-firm demand linkage analyzed by Murphy et al. (1989). Owners, by contrast, spend none of their income in the sector in the baseline (Section 5.4 relaxes this). This MPC asymmetry implies that when automation displaces workers, income shifts toward agents with a lower sectoral MPC, reducing aggregate expenditure on the sector.

When firm j automates a fraction α_j of its tasks, $\alpha_j L$ workers are displaced. A fraction $\eta \in [0, 1]$ of displaced wage income is replaced via reemployment, transfers, or other sources (Jacobson et al., 1993); the remainder, $(1-\eta)w$ per displaced worker, is lost to the sector.

Across all N firms, the total number of displaced workers is $\sum_j \alpha_j L$, so total wage income lost to displacement is $(1-\eta)w \sum_j \alpha_j L$. Total labor income in the sector is therefore $wLN - (1-\eta)w \sum_j \alpha_j L$, of which a fraction λ is spent on the sector's good. Adding autonomous demand $A > 0$ (from outside the sector or from capital income), aggregate sectoral expenditure is

$$D(\boldsymbol{\alpha}) = A + \lambda w L\Big[N - (1-\eta)\textstyle\sum_j \alpha_j\Big]. \tag{2}$$

Writing $\bar{\alpha} := \frac{1}{N}\sum_j \alpha_j$ for the average automation rate, this becomes $D = A + \lambda w L N[1 - (1-\eta)\bar{\alpha}]$. Defining the effective demand loss per automated task as

$$\ell := \lambda(1-\eta)w, \tag{3}$$

this simplifies to $D = A + \lambda w L N - \ell L N \bar{\alpha}$: demand falls linearly in the average automation rate.

Firms sell their output on the product market at a uniform price that equates aggregate supply and demand. Since all firms produce the same output $Y_i = L$, total supply is NL and the market-clearing price is $p = D/(NL)$. Each firm earns revenue $\mathrm{Rev}_i = p \cdot Y_i = D/N$, which, after substituting (2), gives

$$\mathrm{Rev}_i = \frac{A}{N} + \lambda w L - \ell L \bar{\alpha}. \tag{4}$$

Firm i's profit is $\pi_i = \mathrm{Rev}_i - C_i$. Substituting (4) and (1):

$$\pi_i = \Pi_0 + L\bigl(s\,\alpha_i - \ell\bar{\alpha} - \tfrac{k}{2}\alpha_i^2\bigr)\,, \tag{5}$$

where $\Pi_0 \coloneqq A/N + (\lambda - 1)wL$ is the per-firm profit when no firm automates. Writing $\bar{\alpha} = (\alpha_i + \sum_{j\neq i}\alpha_j)/N$ to isolate firm i's own action:

$$\pi_i = \Pi_0 + L\left[\alpha_i\left(s - \frac{\ell}{N}\right) - \frac{k}{2}\alpha_i^2 - \frac{\ell}{N}\sum_{j\neq i}\alpha_j\right]. \tag{6}$$

Firms play a one-shot simultaneous-move game, each choosing α_i to maximize π_i; the product market then clears mechanically given the automation profile.[2] The solution concept is Nash equilibrium.

Define aggregate owner surplus $\mathcal{K}$ and aggregate worker income $\mathcal{W}$:

$$\mathcal{K} \coloneqq \sum_i \pi_i$$

$$\mathcal{W} \coloneqq wLN[1 - (1-\eta)\bar{\alpha}].$$

We measure over-automation against two benchmarks: the *cooperative optimum*, which maximizes $\mathcal{K}$, and a *generalized social planner* who maximizes

$$S(\mu) \;\coloneqq\; \mu\,\mathcal{W} + (1-\mu)\,\mathcal{K}$$

for a weight $\mu \in [0,1]$ on workers.

Note that the environment assumes full transparency: every firm can directly observe how automation maps into lost worker income and reduced aggregate spending. The question Section 3 answers is whether this visibility alone is sufficient for firms to curb automation in a competitive setting.

---

[2] An alternative would be a two-stage game in which firms first choose automation rates and then compete on price or quantity. We abstract from this type of more elaborate second-stage product-market competition because those strategic effects are already well studied and would obscure the novel mechanism we isolate here: the demand externality from automation under full transparency. The qualitative results are plausibly robust to richer product-market interaction, but closed-form solutions would become substantially more complex.

3 Equilibrium and Over-Automation

This section derives the equilibrium, shows firms over-automate relative to the cooperative optimum, and quantifies the resulting surplus loss. All proofs are collected in Section A.

3.1 Equilibrium and the Over-Automation Wedge

To characterize the equilibrium, consider firm i's marginal incentive to automate. Recall from (3) that $\ell(w) = \lambda(1-\eta)w$ is the demand lost per displaced worker, proportional to the wage because displaced workers' forgone spending scales with their earnings. (We write simply ℓ when the wage is held fixed; Section 5.3 endogenizes w.) From (6), firm i's marginal profit from automation is

$$\frac{\partial \pi_i}{\partial \alpha_i} = L\left(s - \frac{\ell}{N} - k\alpha_i\right).$$

A marginal increase in automation saves s in labor costs but incurs friction $k\alpha_i$ and reduces the firm's revenue by ℓ/N. The revenue loss is ℓ/N rather than ℓ because competitive pricing allocates revenue equally across symmetric firms (4): firm i's automation reduces aggregate demand by ℓL, but only $\ell L/N$ of this falls on firm i itself. Each firm therefore underestimates the social cost of its automation, suggesting systematic over-automation in equilibrium. The following proposition confirms this and quantifies the gap.

Proposition 1 (Equilibrium and over-automation)**.** *In the model defined in Section 2, define the automation threshold*

$$N^* := \frac{\ell}{s} = \frac{\lambda(1-\eta)w}{w-c}. \tag{7}$$

If $N \le N^*$, no firm automates ($\alpha^{NE} = 0$).

If $N > N^*$ (equivalently, $s > \ell/N$):

- *(i) Each firm's strictly dominant strategy is $\alpha^{NE} = \min\left((s-\ell/N)/k, 1\right)$;*
- *(ii) The cooperative optimum is $\alpha^{CO} = \min\left(\max\left(0, (s-\ell)/k\right), 1\right)$;*
- *(iii) If $\ell < s < k + \ell/N$ then both α^{NE} and α^{CO} are interior, and the over-automation wedge is*

  $$\alpha^{NE} - \alpha^{CO} = \frac{\ell\,(1 - 1/N)}{k} > 0.$$

  This is strictly increasing in N and ℓ, and decreasing in k.
- *(iv) If $s \le \ell$, then $\alpha^{CO} = 0$, and so the wedge is α^{NE}. Thus if $s < k + \ell/N$, then $\alpha^{NE} = (s - \ell/N)/k$, so the wedge is $(s-\ell/N)/k$. On the other hand, if $k + \ell/N \le s$, then $\alpha^{NE} = 1$,*

so the wedge is 1.

The proposition follows from the private first-order condition derived above and its cooperative counterpart: a planner setting a common rate for all firms faces the full demand loss ℓ per automated task rather than the ℓ/N each firm perceives, yielding $\alpha^{CO} = (s - \ell)/k$. Because rivals' rates enter (6) only through the term $-(\ell/N)\sum_{j\neq i}\alpha_j$, which is independent of α_i, the equilibrium rate is a strictly dominant strategy: each firm over-automates even with perfect foresight about every rival's behavior.

The case structure arises because both α^{NE} and α^{CO} lie in $[0, 1]$: each can be at no automation, interior, or full automation depending on how the cost saving s compares to the demand-loss and friction parameters. The over-automation wedge is largest when the cooperative rate is zero but individual firms still find automation privately worthwhile, and it vanishes when both rates hit the same boundary. Parts (iii)–(iv) enumerate the relevant combinations; the economic force is the same throughout.

The wedge is strictly increasing in N: more competitive sectors exhibit wider automation gaps. This runs counter to the standard intuition that competition disciplines firms to act in consumers' interests; here, more competition dilutes each firm's share of the demand loss, weakening the private incentive to restrain. A monopolist ($N = 1$) fully internalizes the externality ($\alpha^{NE} = \alpha^{CO}$); as $N \to \infty$, the wedge approaches its maximum of ℓ/k.

From Proposition 1, a firm automates only when $N > N^* = \ell/s$: the number of competitors must be large enough that each firm's share of the demand loss, ℓ/N, falls below its cost saving s. As AI costs fall ($c \to 0$), $N^* \to \lambda(1-\eta) \leq 1$: the over-automation region expands to cover virtually any market with $N \geq 2$. For illustrative parameters ($c/w = 0.30$, $\lambda = 0.5$, $\eta = 0.30$, $N \to \infty$), the wedge equals $\ell/k = \alpha^{CO}$: firms in competitive markets automate at twice the cooperatively efficient rate.

Figure 1 illustrates these comparative statics. In each panel, the dashed line marks the $N = N^*$ boundary below which no firm automates, and darker shading indicates a larger wedge. The dominant pattern is that the wedge grows with N; non-monotonicity in the other dimensions reflects the regime shift at $s = \ell$, where the cooperative optimum moves from zero to an interior solution.

When frictions are positive, adjustment costs moderate the equilibrium automation rate. The next subsection shows that when frictions vanish ($k \to 0$), this moderating force disappears and the game reduces to a Prisoner's Dilemma: full automation versus none.

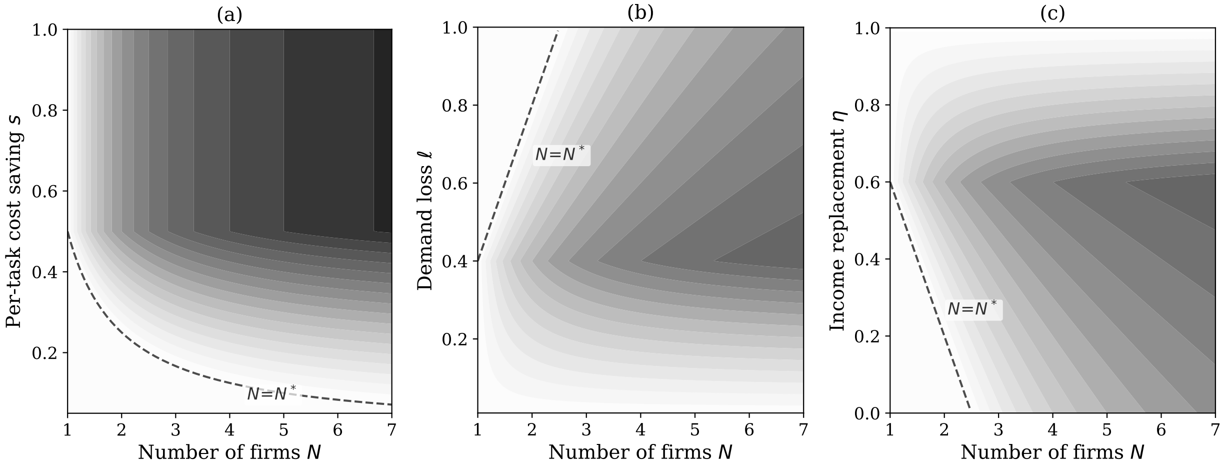

Figure 1: The over-automation wedge $\alpha^{NE} - \alpha^{CO}$ across the parameter space. Shading runs from white (zero wedge) to black (wedge ≥ 0.50). Dashed lines mark the $N = N^*$ boundary below which no firm automates. (a) Number of firms vs. cost saving $s = w - c$ at fixed wage ($\lambda = 0.5$, $\eta = 0$, $k = 1$). (b) Number of firms vs. demand loss $\ell = \lambda(1 - \eta)w$ ($c = 0.60$, $\eta = 0$, $k = 1$). (c) Number of firms vs. income replacement ($c = 0.60$, $\lambda = 1$, $k = 1$).

3.2 Frictionless Automation as a Prisoner's Dilemma

When $k = 0$, marginal profit becomes the constant $L(s - \ell/N)$, independent of the automation level, and the outcome is all-or-nothing. If $N \leq N^*$, no firm automates. If $N > N^*$, automating is strictly dominant yet collectively harmful:

Corollary 1 (Frictionless limit)**.** *Suppose adjustment frictions vanish* ($k = 0$) *and the number of firms exceeds the critical threshold* ($N > N^*$)*.*

(i) Full automation ($\alpha_i = 1$) *is strictly dominant for every firm.*

(ii) If additionally the cost saving is less than the demand loss per task ($s < \ell$)*, the cooperative optimum is no automation* ($\alpha_i = 0$ *for all* i*, yielding per-firm profit* Π_0)*; the equilibrium yields* $\Pi_0 + L(s - \ell) < \Pi_0$*. Total deadweight loss is* $NL(\ell - s)$*.*

Under the condition in part (ii) ($s < \ell$), the Prisoner's Dilemma structure makes the failure of voluntary restraint transparent. A firm that holds back unilaterally (choosing $\alpha_i = 0$) still suffers the revenue decline from rivals' automation but forgoes the offsetting cost savings; a firm that deviates (choosing $\alpha_i = 1$) captures the savings while imposing only a $1/N$ share of the demand loss on itself. The resulting payoff matrix has the classic form: mutual restraint yields Π_0 per firm, while mutual automation yields $\Pi_0 + L(s - \ell) < \Pi_0$, yet defecting is individually rational regardless of others' choices. Because automating is strictly dominant (not merely a best response to others' automating), no non-binding agreement can restore efficiency. Communication is cheap talk in the sense of Crawford and Sobel (1982): even if all firms acknowledge that collective restraint would raise profits, each firm's individually optimal action remains unchanged. This distinguishes the automation externality from pure coordination failures (where firms simply need to agree on which equilibrium to play) and motivates the analysis of Coasian bargaining in Section 4.5.

3.3 Over-Automation as Deadweight Loss

Is the over-automation wedge merely a redistribution from workers to firm owners, or does it reduce total surplus? Recall the generalized planner introduced in the model section, who maximizes

$$S(\mu) \;=\; \mu\,\mathcal{W} + (1-\mu)\,\mathcal{K} \tag{8}$$

for a weight $\mu \in [0,1]$ on workers.

Proposition 2 (Generalized planner and surplus loss). *Suppose* $k > 0$ *and* $N > N^*$.

(i) The μ*-planner's optimal automation rate is*

$$\alpha^{SP}(\mu) := \frac{s-\ell}{k} - \frac{\mu\,\ell}{\lambda(1-\mu)\,k},$$

where, as usual, $\alpha^{SP}(\mu)$ *is the automation rate, and is thus restricted to the interval* $[0,1]$*. At* $\mu = 0$ *this reduces to* α^{CO} *from Proposition 1.*

(ii) The surplus loss from the Nash equilibrium relative to the planner's optimum is

$$S\big(\mu;\alpha^{SP}\big) - S\big(\mu;\alpha^{NE}\big) = \frac{(1-\mu)\,NLk}{2}\,\big[\alpha^{NE} - \alpha^{SP}(\mu)\big]^2.$$

(iii) (Pareto dominance.) $\alpha^{NE} > \alpha^{SP}(\mu)$ *for every* $\mu \in [0,1)$*. The Nash equilibrium is Pareto dominated by the cooperative optimum: workers and firm owners are both strictly worse off.*

Over-automation is not a transfer from workers to owners: it is a deadweight loss that harms both sides (part (iii)). Workers lose wage income directly through displacement. Firm owners, despite cutting costs on each automated task, also lose: collective displacement erodes demand to the point where every firm's equilibrium profit falls below its cooperative-optimum profit. No redistribution between the two groups can make the Nash outcome efficient.

When both α^{NE} and α^{SP} are interior, the total wedge between equilibrium and the planner's optimum decomposes into two distinct sources:

$$\underbrace{\alpha^{NE} - \alpha^{SP}(\mu)}_{\text{total wedge}} = \underbrace{\frac{\ell(1-1/N)}{k}}_{\text{demand externality}} + \underbrace{\frac{\mu\,\ell}{\lambda(1-\mu)\,k}}_{\text{distributional}}. \tag{9}$$

The first term is the uninternalized demand externality from Proposition 1(iii): it is present even when the planner places zero weight on workers ($\mu = 0$) and cares only about aggregate profit. It grows with N, approaching ℓ/k as $N \to \infty$, so fragmented markets suffer disproportionately. The second term is a distributional premium: the additional automation reduction a planner who values worker income ($\mu > 0$) would impose beyond the profit-maximizing benchmark. It is independent of N but grows without bound as $\mu \to 1$; at $\bar{\mu} := \lambda k\,\alpha^{CO}/(\ell + \lambda k\,\alpha^{CO})$ the planner prohibits automation entirely. The surplus loss in (ii) is quadratic in this total wedge and scales with NL, so both fragmentation and market size amplify the welfare cost.

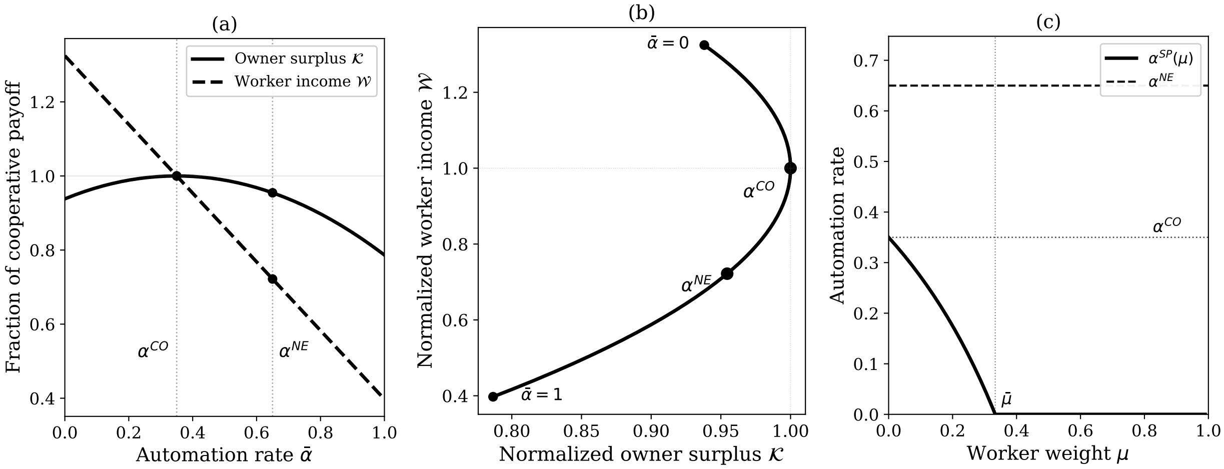

Figure 2: Welfare consequences of over-automation. All panels use $c = 0.30$, $\lambda = 0.5$, $\eta = 0.30$, $N = 7$, $k = 1$, $A = 10$. (a) Normalized owner surplus $\mathcal{K}$ and worker income $\mathcal{W}$ vs. automation rate ($1 =$ cooperative benchmark): both groups lose at α^{NE}. (b) Factor payoff frontier: α^{CO} at $(1, 1)$, α^{NE} strictly southwest (Pareto dominated). (c) Planner's optimum $\alpha^{SP}(\mu)$ vs. worker weight μ (not normalized): the gap to α^{NE} is the over-automation wedge, positive even at $\mu = 0$.

Figure 2 illustrates the Pareto dominance and the decomposition. To make the losses for the two groups comparable, panels (a) and (b) normalize each payoff by its value at α^{CO}, so a value of 1 corresponds to the cooperative benchmark. Panel (a) plots the normalized payoffs against the common automation rate $\bar{\alpha}$: both curves peak at or before α^{CO} and both fall below 1 at α^{NE}. The key observation is that the equilibrium rate lies to the right of the aggregate profit peak, so that both owner surplus and worker income are lower than under cooperation. Workers bear the larger loss because their income declines linearly in $\bar{\alpha}$, while the profit curve is concave and falls more gently. Panel (b) re-expresses the same information as a factor payoff frontier: each point on the curve corresponds to a different common automation rate, tracing out the $(\mathcal{K}, \mathcal{W})$ pairs as $\bar{\alpha}$ increases. The cooperative rate sits at $(1, 1)$ and α^{NE} is strictly to the southwest, confirming that moving from equilibrium to the cooperative rate would make both groups better off. Panel (c) visualizes the decomposition in (9): the horizontal line marks α^{NE}, and the declining curve is the planner's optimum $\alpha^{SP}(\mu)$. Even at $\mu = 0$, the gap is positive (the demand-externality term alone), and the required correction grows further as the distributional premium widens with μ.

Since the over-automation wedge is a structural externality that harms both factor classes, a natural question is whether policy can close it.

4 Policy Instruments

Several instruments could in principle address the externality; the question is which ones operate on the right margin. To answer it, we benchmark against the cooperative optimum α^{CO}, which maximizes aggregate profit without directly weighting worker welfare. This is deliberately the weakest case for intervention: Proposition 2 shows that the demand externality alone reduces both firm profits and worker income, and that placing any positive weight on workers ($\mu > 0$) only widens the wedge. Adopting the efficiency benchmark therefore simplifies the exposition while providing the most conservative case for intervention.

Table 1 previews the results: only the Pigouvian automation tax fully corrects the distortion; the remaining instruments cushion the losers or partially shrink the wedge, but none eliminates it.

Table 1: Policy instruments and their effects on the over-automation externality.

	Section	Changes N^*?	Changes wedge?	Fixes externality?
Upskilling/retraining (η)	4.1	Yes	Yes	Partially
Universal Basic Income (A)	4.2	No	No	No
Capital income tax (t)	4.3	No	No	No
Worker equity (ϵ)	4.4	Yes	Yes	Partially
Coasian bargaining ($M < N$)	4.5	No	Partially	No
Automation tax (τ)	4.6	Yes	Yes	**Yes**

Note. Each row records whether the instrument alters the automation threshold N^*, the over-automation wedge $\ell(1 - 1/N)/k$, and whether it fully corrects the demand externality.

One limitation should be noted: the analysis evaluates each instrument against a single margin, the demand externality identified in Section 3, holding all other features of the economy fixed. In practice, every instrument carries additional costs and benefits outside the model (administrative burden, labor-market distortions, political feasibility) that a full welfare analysis would need to weigh. Nonetheless, an instrument that does not operate on the externality margin cannot correct the distortion regardless of how it scores on other dimensions; the analysis below separates instruments that can from those that cannot.

4.1 Displacement vs. Upskilling

The demand-loss parameter $\ell = \lambda(1-\eta)w$ governs the externality's magnitude. In the baseline model, $\eta \in [0,1]$ represents the fraction of displaced wage income recovered through reemployment, transfers, or other sources: higher η shrinks ℓ and thereby the over-automation wedge.

But the parameter extends naturally beyond unity. When $\eta > 1$, upskilling and reabsorption place displaced workers into higher-paying roles, automation *increases* aggregate labor income, and loss ℓ turns negative, which we can interpret as a gain. This is the scenario invoked by AI optimists, in which technological displacement is a stepping stone to better jobs. As the following corollary shows, the sign reversal in ℓ flips the externality itself.

Corollary 2 (Sign of the externality)**.** *For any $\eta \geq 0$ and $s > \ell$, the over-automation wedge is $(\alpha^{NE} - \alpha^{CO}) = \ell\,(1 - 1/N)/k$, which is maximized at $\eta = 0$ (where $\ell = \lambda w$), positive for all $\eta < 1$, zero when $\eta = 1$ (where $\alpha^{NE} = \alpha^{CO} = s/k$), and negative (under-automation) when $\eta > 1$.*

The logic is symmetric. When $\eta < 1$, displacement destroys demand, and each firm bears only $1/N$ of the loss, producing over-automation. When $\eta > 1$, displacement *creates* demand through higher reemployment wages, and each firm captures only $1/N$ of the gain, producing under-automation. In both cases, the distortion grows with N: more competition dilutes each firm's share of the externality, whether that externality is negative or positive. A monopolist ($N = 1$) fully internalizes in every case.

The competitive forces are identical; only the sign of the demand externality differs. As Section 4.6 will show, the same corrective instrument addresses both cases: a tax when $\eta < 1$, a subsidy when $\eta > 1$.

The case $\eta > 1$ is not merely theoretical. Historical technological transitions have often eventually reabsorbed displaced workers at higher wages (Acemoglu and Restrepo, 2019), and the current AI buildout offers a concrete channel: the expansion of data centers, energy infrastructure, and AI-adjacent services is creating skilled roles that can pay more than the positions automation displaces. If this reabsorption is fast enough to push η above unity, competitive firms will automate *too slowly*. However, past displacement episodes have consistently produced $\eta < 1$: displaced workers suffer large, persistent earnings losses (Jacobson et al., 1993), and there is little evidence yet that AI-driven displacement will differ, placing most economies firmly in the over-automation regime.

The policy implication is that raising η through retraining programs, wage insurance, and incentives for new firm creation is not merely a palliative for displaced workers but a direct lever on the externality: every unit increase in η toward unity shrinks ℓ, narrows the over-automation wedge, and reduces the burden placed on the corrective instruments analyzed below. Pushing η past unity would flip the distortion into under-automation, but this is a far less pressing concern: in that regime, displaced workers are already thriving in higher-paying roles.

4.2 Universal Basic Income

Among the most discussed responses to automation-driven displacement is a universal basic income. In the model, a UBI funded from general revenue maps to an increase in autonomous demand A: because the transfer is unconditional, employed and displaced workers receive the same payment, adding a constant to aggregate spending without altering the marginal income loss from displacement. This distinguishes UBI from displacement-targeted transfers (wage insurance, severance), which raise the income-replacement rate η and directly shrink ℓ; see Section 4.1. The results below concern this modeled object and should not be read as a verdict on all UBI designs.

Because UBI adds a constant to demand, it enters firm profit only through $\Pi_0 = A/N + (\lambda - 1)wL$, the baseline profit when no firm automates. This term drops out of the first-order condition $s - \ell/N - k\alpha_i = 0$: a higher A raises the profit floor but changes neither the cost saving s nor the demand loss ℓ that determine the automation rate. Consequently, UBI alters neither the automation threshold $N^* = \ell/s$ nor the over-automation wedge $\ell(1 - 1/N)/k$. In the language of game theory, UBI changes payoff *levels* but not the payoff *differences* that drive strategic behavior. More generally, instruments that operate on profit levels can redistribute income but cannot correct the externality; only instruments that change the per-task automation margin can.

Despite not correcting the externality, UBI serves a complementary role. A higher Π_0 cushions profit losses from over-automation, while the transfer itself raises the floor on workers' living standards, buying time for corrective instruments that operate on the right margin.

UBI may, however, carry an unintended side effect when the number of firms is endogenous. The baseline fixes N, but if firms can freely enter (Section 5.2 formalizes this), higher profits attract new entrants, fragmenting the market. Because the over-automation wedge is increasing in N, UBI-induced entry can paradoxically widen the externality, partially offsetting the welfare gain from higher baseline consumption; Section 5.2 develops the full argument.

Within the model, UBI is a *complement* to the automation tax, not a substitute: a society that relies solely on UBI will over-automate at the same rate, with a higher floor on living standards but the same externality.

4.3 Capital Income Taxation

If unconditional transfers do not alter the automation incentive, a natural alternative is to tax the *proceeds* of automation directly. Consider a proportional tax $t \in (0, 1)$ on capital income (profits), with revenue redistributed to workers. Firm i now maximizes $(1 - t)\pi_i$, but because $(1 - t)$ is a positive scalar it cancels from the first-order condition: the equilibrium automation rate, the threshold N^*, and the over-automation wedge are all unchanged.

The revenue side fares no better. Lump-sum redistribution raises autonomous demand A, but A enters per-firm profit only through the constant $\Pi_0 = A/N + (\lambda - 1)wL$, which does not appear in the first-order condition. If revenue instead funds displacement insurance that raises η, the externality shrinks through ℓ, but the operative channel is η, not the profit tax itself.

The distinction matters because capital income taxes are often conflated with robot taxes in the policy debate. The robot taxes studied in the literature (e.g., Guerreiro et al., 2022) are per-unit levies on adoption, which operate on the per-task margin; a proportional capital income tax is a fundamentally different instrument that scales the entire profit function by $(1 - t)$ and cancels from the optimality condition. The failure is structurally identical to that of UBI (Section 4.2): both instruments shift profit levels rather than operating on the margin where the externality resides.

4.4 Worker Equity Participation

A market-based alternative to taxation, rooted in the profit-sharing literature (Weitzman, 1985), gives workers a direct stake in the profits that automation generates. Unlike UBI, which enters demand only through the level term A, profit-sharing flows through the profit function and therefore interacts with automation decisions. Suppose each firm distributes a fraction $\epsilon \in [0, 1]$ of its profits to workers (through ESOPs, equity grants, or co-determination mandates). Workers spend a λ-fraction of this income in the sector, so profit-sharing recycles capital income back into demand.

Aggregate demand now satisfies a fixed-point condition: $D = A + \lambda[\text{wage income} + \epsilon \sum_i \pi_i]$, where $\sum_i \pi_i = D - \sum_i C_i$. Because profits depend on D, demand is determined simultaneously with the automation decision; the proof solves this fixed point explicitly.

Proposition 3 (Worker equity reduces but cannot eliminate the wedge)**.** *Let* $\epsilon \in [0,1]$ *and* $k > 0$*, and suppose the equilibrium is interior. Define* $N_\epsilon := N - \lambda\epsilon(N-1)$*.*

(i) The cooperative optimum is unchanged: $\alpha^{CO}(\epsilon) = (s-\ell)/k$*, where, as usual* α^{CO} *is restricted to* $[0,1]$*.*

(ii) The Nash equilibrium automation rate is $\alpha^{NE}(\epsilon) = (s - \ell/N_\epsilon)/k$ *(restricted to* $[0,1]$*).*

(iii) When both α^{NE} *and* α^{CO} *are interior solutions, the over-automation wedge*

$$\alpha^{NE}(\epsilon) - \alpha^{CO} = \frac{\ell(N-1)(1-\lambda\epsilon)}{k\,N_\epsilon}$$

is strictly decreasing in ϵ *but strictly positive for all* $\epsilon < 1/\lambda$*. The wedge vanishes only at* $\epsilon = 1/\lambda$*, which requires* $\epsilon > 1$ *whenever* $\lambda < 1$*.*

Part (i) is not immediate: profit-sharing changes the demand function by recycling capital income into worker spending, so one might expect it to shift the planner's optimum. The result follows because the planner already controls all N firms and thus fully internalizes the demand externality. In the planner's first-order condition, the profit-sharing terms cancel: the modified demand-loss parameter $\ell_\epsilon = \ell - \lambda\epsilon s$ and the demand multiplier $1/(1-\lambda\epsilon)$ exactly offset, leaving $k\alpha = s - \ell$ regardless of ϵ.

While the cooperative optimum is unaffected, the Nash equilibrium does shift. Intuitively, when workers hold equity, part of the demand lost through displacement is recycled back through profit shares; each firm therefore perceives a larger effective demand loss from its own automation than in the baseline, and restrains accordingly. The magnitude of this shift is governed by the compound parameter $N_\epsilon = N - \lambda\epsilon(N-1)$. This parameter measures the effective demand-leakage divisor: at $\epsilon = 0$ it equals N, recovering the baseline in which each firm perceives demand loss ℓ/N per automated task. As ϵ rises, N_ϵ falls toward 1, pushing α^{NE} toward the cooperative optimum.

Despite this improvement, the recycling cannot fully close the wedge whenever $\lambda < 1$. Closing the wedge requires the product $\lambda\epsilon$ to reach one, i.e., $\epsilon = 1/\lambda$. When $\lambda < 1$ this exceeds the feasible range $\epsilon \in [0,1]$: each unit of profit recycled to workers generates only λ units of sectoral demand, so compensating for the leakage would require sharing more than the firm's entire profit. Even at $\epsilon = 1$ (full profit-sharing), the wedge reduces to $\ell(N-1)(1-\lambda)/[k(N-\lambda(N-1))]$, which remains strictly positive. (The knife-edge case $\lambda = 1$ is the exception: full profit-sharing does close the wedge, but only because this extreme assumption eliminates the very spending leakage that drives the externality.)

The structural limitation is that the externality is fundamentally multilateral: each firm's automation depresses demand for all N firms, and bilateral arrangements between a firm and its own workers cannot reach the demand that leaks to rivals.

A separate question is whether profit-sharing would arise voluntarily.

Corollary 3 (Voluntary profit-sharing does not arise)**.** *If each firm independently chooses its own profit-sharing rate* $\epsilon_i \in [0, 1]$ *to maximize retained profit* $(1-\epsilon_i)\pi_i$*, then* $\epsilon_i = 0$ *is a dominant strategy.*

The marginal cost of sharing is π_i (a dollar-for-dollar reduction in retained earnings), while the marginal demand benefit is only $\lambda\pi_i/N$: workers spend fraction λ of the shared profit in the sector, and firm i captures $1/N$ of the resulting demand increase. Since $\lambda/N < 1$ for any $N \geq 2$, the cost strictly exceeds the benefit. This is a second-order coordination failure layered on top of the automation externality itself, mirroring the Prisoner's Dilemma structure of Section 3.2.

Profit-sharing must therefore be mandated to have any effect, and even then it cannot substitute for a corrective tax: it narrows the wedge but cannot eliminate it, and unlike a corrective tax (Section 4.6), does not generate government revenue for retraining programs that would raise η.

4.5 Coasian Bargaining

None of the instruments considered so far fully closes the over-automation wedge, and worker equity will not arise voluntarily (Corollary 3). A natural question is whether private ordering could succeed where these instruments have not. By the Coase Theorem (Coase, 1960), if property rights over the externality were well-defined and transaction costs sufficiently low, bargaining could achieve the cooperative optimum without government intervention. A contemporary version of this argument envisions each worker equipped with an AI agent that bargains on her behalf, drawing on the property rights over training data proposed by Arrieta-Ibarra et al. (2018) and Posner and Weyl (2018). To evaluate this possibility, it helps to separate two questions: can bargaining between a firm and its own workers correct the externality, and can bargaining among firms do so? As we show below, neither can.

Bargaining between a firm and its own workers. If displaced workers can bargain for compensation, say a per-task severance payment σ, the firm's effective cost saving falls from s to $s - \sigma$, and the equilibrium automation rate drops. But the severance also recycles income into demand: displaced workers spend a fraction λ of their compensation in the sector, reducing the effective demand-loss parameter from ℓ to $\ell - \lambda\sigma$. The over-automation wedge therefore becomes $(1 - 1/N)(\ell - \lambda\sigma)/k$, which is smaller but still positive. This is operationally equivalent to raising the income-replacement rate η by σ/w; the same logic applies to equity stakes through the profit-sharing parameter ϵ (analyzed in Section 4.4). As Section 4.1 shows, raising η narrows the wedge but cannot close it: the externality persists as long as displaced workers' spending is not fully replaced, which requires $\lambda\sigma = \ell$, i.e., full income replacement through bargaining alone.

Moreover, the uninternalized portion of the externality does not fall on firm i's own workers at all. When firm i automates, the demand loss $\ell(1 - 1/N)L$ reduces revenue at rival firms, lowering their owners' profits (Equation (4)). Workers at rival firms who retain their positions continue to earn w per task; they are not directly harmed and have no basis for negotiation with firm i. The externality is therefore a firm-to-firm channel running through the product market, not a firm-to-worker channel that bilateral bargaining can reach.

Firm-to-firm bargaining. Since the externality runs across firms, consider a coalition of $M \leq N$ firms that jointly choose automation rates to maximize their combined profit, while the remaining $N - M$ firms play Nash.

Proposition 4 (Partial coalitions cannot eliminate the wedge)**.** *Let $k > 0$ and suppose the equilibrium is interior. A coalition of M firms that jointly maximizes its members' combined profit chooses the common automation rate*

$$\alpha^M = \frac{s - M\ell/N}{k}.$$

The residual over-automation wedge relative to the cooperative optimum is

$$\alpha^M - \alpha^{CO} = \frac{\ell(1 - M/N)}{k} > 0 \quad \textit{for all } M < N.$$

The wedge vanishes only when $M = N$: only the grand coalition replicates the cooperative optimum.

The intuition is that a coalition of M firms internalizes M/N of the aggregate demand loss, but the remaining fraction $(1 - M/N)$ is still externalized onto non-members. Four features of the automation externality prevent the grand coalition from forming. First, voluntary agreements are not self-enforcing: in the frictionless limit (Corollary 1), automation is strictly dominant, so a coalition member gains from deviating regardless of whether others honor the agreement. This is not a coordination failure that communication can resolve; the Prisoner's Dilemma structure means no non-binding arrangement is stable. With convex costs ($k > 0$), the deviation incentive is continuous but still positive. Second, the externality is multilateral and diffuse. The Coase Theorem's canonical applications involve bilateral or small-number settings; here, each of N firms imposes demand losses on all $N - 1$ others. Each firm's individual contribution to the demand loss is $\ell L/N$, too small to motivate any single negotiation yet too large in aggregate to ignore. This is precisely the large-numbers setting in which Coase (1960) himself acknowledged that private bargaining breaks down. Third, the automation rate α_i is not contractible among firms: it is an internal organizational choice that rival firms cannot observe or verify, making binding private agreements impractical. Fourth, automation decisions involve large sunk costs and are substantially irreversible, so even in a repeated setting, trigger-strategy punishments cannot undo a deviation; a firm that delays while rivals proceed loses market share (see Section 5.1 below); and large N makes defection harder to detect and punishment harder to sustain.

To summarize, the demand externality studied here is not a market failure that private ordering can cure. Worker-side bargaining operates on within-firm channels (η, ϵ) that cannot reach the cross-firm margin where the externality resides; firm-to-firm bargaining targets the right margin but cannot sustain the grand coalition needed to close the wedge. The fundamental obstacle is incentive compatibility, not transaction costs: even with costless negotiation, the automation game retains its dominant-strategy structure. Correcting the externality therefore requires an instrument that does not rely on voluntary agreement but instead changes each firm's marginal automation incentive directly.

4.6 Pigouvian Automation Tax

The classic remedy for a negative externality is a *Pigouvian tax*: a per-unit charge set equal to the marginal external cost, so that every agent's private incentive aligns with the social cost (Pigou, 1920). In contrast to many textbook externalities, where the harmed parties are outside the firms' market (e.g., pollution), here the harmed parties are workers whose income constitutes the firms' own demand. This means the tax rate, its revenue, and its incidence all interact through the same labor-market channel, creating richer policy design questions than the standard case.

Proposition 5 (Pigouvian automation tax)**.** *Let $\tau \geq 0$ be a per-task automation tax, $k > 0$, and suppose $N > N^*$ and the equilibrium is interior.*

(i) The Nash equilibrium automation rate is $\alpha^{NE}(\tau) = (s - \tau - \ell/N)/k$. The rate

$$\tau^* = \ell\left(1 - \frac{1}{N}\right)$$

implements $\alpha^{NE} = \alpha^{CO} = (s - \ell)/k$.

(ii) Under the tax at rate τ^* without rebate, each firm earns $\pi^{\text{tax}} = \pi^{CO} - \tau^* L\alpha^{CO}$. With lump-sum rebate of tax revenue, each firm achieves exactly π^{CO}.

The optimal rate has a transparent economic interpretation: each firm already bears ℓ/N of the demand loss from its own automation; the tax charges it for the remaining $\ell(1-1/N)$ imposed on rivals. For large N, $\tau^* \approx \ell = \lambda(1-\eta)w$, so setting the rate requires only sector-level observables. Levying the tax, however, requires observing firm-level automation rates, a practical challenge, though one that may be easing as AI adoption generates observable procurement records (Guerreiro et al., 2022). Unlike rival firms in a Coasian bargaining setting (Section 4.5), a tax authority can compel disclosure through mandatory reporting, payroll records, and procurement audits, making approximate measurement feasible even when private verification is not. Because the welfare loss is quadratic in the wedge (Proposition 2), even an imprecisely targeted tax yields a first-order gain.

Allocation of tax revenue. Proposition 5 pins down the rate; the remaining design question is what to do with the revenue. Because the externality flows through the labor market, this choice can affect structural parameters governing the distortion, not merely the distribution of gains.

A lump-sum rebate to firms restores cooperative profits exactly (part (ii)), but returns revenue to the firms that automate while leaving displaced workers, the harmed parties, uncompensated. The more natural option is to direct revenue toward those workers. Two channels are available, with different incentive properties.

Direct transfers (wage insurance, severance supplements) raise η mechanically by replacing lost income. The firm's automation incentive is unaffected: it pays τ per automated task regardless of where the revenue goes, so its first-order condition does not change. But the standard moral-hazard concern applies: generous income replacement may weaken workers' incentive to retrain or reallocate, sustaining η through transfers rather than through productive reabsorption.

Funding retraining programs also raises η, but through human-capital investment rather than income replacement. This channel is slower and harder to implement, yet it builds the capacity for workers to re-enter the labor market at comparable or higher wages, making gains in η self-sustaining. In principle, the resulting dynamic is self-reinforcing, complementing the analysis in Section 4.1: the tax funds programs that raise η, which lowers ℓ, which reduces τ^* in future periods. To the extent that reabsorption is successful, the required correction shrinks over time and the tax is transitional rather than permanent, echoing the finding in Guerreiro et al. (2022) that the optimal robot tax declines to zero as displaced cohorts retire and new workers choose occupations with full knowledge of automation.

In practice, a mix of the two is likely optimal: short-run transfers to cushion displacement while longer-run retraining builds durable gains in η. The Pigouvian tax therefore has the potential to do double duty: it corrects the externality at the margin, and its revenue can be recycled to shrink the distortion over time.

5 Extensions

The baseline isolates the demand externality in the simplest environment that supports it. A natural concern is that the result depends on what has been held fixed: endogenous wage adjustment might close the wedge, free entry might discipline the market to an efficient scale, higher AI productivity might resolve the demand problem by expanding the pie, and capital-income recycling might offset the spending lost through displacement. This section takes up each of these objections, along with richer product-market interaction, and shows that the externality is robust to all of them and, in some cases, amplified. Table 2 previews the results.

Table 2: Extensions of the baseline model and effects on the over-automation result.

	Section	Modifies N^*?	Wedge effect	Eliminates externality?
AI productivity ($\phi > 1$)	5.1	No	Widens	No
Endogenous entry	5.2	No	Persists (can widen)	No
Endogenous wages	5.3	Yes (raises)	Narrows	No
Capital-income recycling ($\hat{\eta} > 0$)	5.4	Yes (raises)	Narrows	Partially
Imperfect competition	5.5	—	Persists	No

Note. Each row records how the generalization affects the automation threshold N^*, the over-automation wedge, and whether it eliminates the demand externality. The externality survives all five generalizations under empirically plausible parameters.

5.1 AI Productivity

In the baseline, AI and human workers produce the same output per task and so the automation incentive is purely cost-driven. In practice, AI can replace humans while also raising output per task (e.g., autonomous agentic coding agents, higher-throughput customer service bots). To capture this, we add a productivity advantage on top of the cost saving. A natural conjecture is that this output channel mitigates the demand problem by making the economy more productive. We show the opposite is true: higher AI output per task widens the over-automation wedge.

Let an AI-performed task produce $\phi \geq 1$ units of output, while a human-performed task produces 1 unit. With $\phi > 1$, firm i's output becomes

$$Y_i(\alpha_i) = [\phi\alpha_i + (1 - \alpha_i)]L = [1 + (\phi - 1)\alpha_i]L.$$

Under perfect competition, revenue is allocated by output share: $\text{Rev}_i = D \cdot Y_i / \sum_j Y_j$. At a symmetric profile all firms produce the same output $\bar{Y} = [1 + (\phi - 1)\alpha]L$, so $\mathrm{Rev}_i = D/N$ as in the baseline. Differentiating with respect to α_i and evaluating at the symmetric profile yields

$$\left.\frac{\partial \,\mathrm{Rev}_i}{\partial \alpha_i}\right|_{\mathrm{sym}} = \underbrace{-\frac{\ell L}{N}}_{\text{demand loss}} + \underbrace{\frac{D(\phi - 1)(N - 1)}{N^2[1 + (\phi - 1)\alpha]}}_{\text{market-share gain}}. \tag{10}$$

The first term is the baseline demand externality, which depends only on ℓ and N and is therefore independent of ϕ. The second is new: a deviating firm raises its output above rivals and captures a larger share of expenditure. This market-share gain is positive whenever $\phi > 1$, raising the private incentive to automate above the baseline.

To quantify the effect, we combine cost saving, demand loss, and market-share gain into the first-order condition. The symmetric equilibrium equates marginal integration cost to the combined benefit:

$$k\alpha = s - \frac{\ell}{N} + \frac{D(\alpha)\,(\phi - 1)(N - 1)}{N^2[1 + (\phi - 1)\alpha]\,L}.$$

Because $D(\alpha)$ is linear in α, clearing the denominator yields a quadratic whose positive root is the unique equilibrium; however, the resulting expression is less transparent than the baseline formula, so the comparative statics below are established via a monotone crossing argument.

Proposition 6 (AI productivity widens the over-automation wedge)**.** *Let $k > 0$ and suppose the equilibrium is interior.*

- *(i) The Nash equilibrium automation rate is increasing in AI productivity: $\alpha^{NE}(\phi) > \alpha^{NE}(1)$ for $\phi > 1$.*
- *(ii) The cooperative optimum and the generalized planner's optimum, however, are unchanged: $\alpha^{CO}(\phi) = \alpha^{CO}(1)$ and $\alpha^{SP}(\mu; \phi) = \alpha^{SP}(\mu; 1)$ for all $\mu \in [0, 1]$.*
- *(iii) Consequently, the over-automation wedge $\alpha^{NE}(\phi) - \alpha^{SP}(\mu; \phi)$ is strictly increasing in ϕ for every $\mu \in [0, 1]$.*

The mechanism is a Red Queen effect: each firm perceives a market-share gain from automating beyond rivals, but at the symmetric equilibrium all firms expand equally, so the gains cancel. By contrast, the cost saving s enters each firm's profit identically regardless of rivals' choices, so it shifts α^{NE} and α^{CO} equally and leaves the wedge unchanged.

Part (ii) holds because total sectoral revenue equals total expenditure D under market clearing, and D (2) depends on worker income, not output: higher ϕ raises output but lowers the price in proportion, leaving the planner's objective invariant to ϕ. Together with part (i), this yields part (iii): better AI raises the equilibrium automation rate without shifting the efficient benchmark, so the distortion grows with AI capability.

The wider wedge also carries a policy implication. Because the market-share motive adds a second distortion on top of the demand externality, the baseline Pigouvian rate $\tau^* = \ell(1 - 1/N)$ no longer suffices: implementing α^{CO} requires an additional correction equal to the market-share term in (10) evaluated at the cooperative rate, which is strictly positive whenever $\phi > 1$. That said, the lower price means each dollar of spending buys more physical output, so the welfare measure $S(\mu)$, built from nominal flows, understates the real consumption gains from higher ϕ. The proposition identifies a strategic distortion, not a claim that higher AI productivity reduces total welfare on net.

5.2 Endogenous Entry

So far the number of firms has been exogenous. With free entry, one might expect the over-automation problem to be self-correcting: surplus erosion lowers profits, marginal firms exit, and the remaining industry settles at an efficient scale. Whether this logic goes through, however, depends on how the entry margin interacts with the automation decision.

Consider a two-stage game: firms pay a fixed cost $\kappa \geq 0$ to enter, then simultaneously choose automation rates. Given N entrants, the stage-2 Nash equilibrium yields per-firm operating profit $\Pi^*(N)$. A pure-strategy *free-entry equilibrium* is an integer $N \geq 1$ such that

$$\Pi^*(N) \geq \kappa \qquad \text{and} \qquad \Pi^*(N+1) \leq \kappa : \tag{11}$$

incumbents weakly prefer to remain active, while an additional entrant would not recover the fixed cost.

Because the frictionless ($k = 0$) and convex-cost ($k > 0$) regimes shape the profit schedule in qualitatively different ways, we treat each in turn. The frictionless case sets $\lambda = 1$ for clean closed forms; the qualitative results hold for any $\lambda \in (0, 1]$, as the assumption affects only profit levels, not the structure of the entry regimes.

In the frictionless benchmark ($k = 0$), automation is all-or-nothing. When $\ell > s$, the profit schedule drops discretely at N^*: below the threshold no firm automates, while above it full automation is dominant (Corollary 1) and per-firm profit falls by $\Delta := L(\ell - s) > 0$ (Figure 4 in the appendix illustrates). Write $m := \lfloor N^* \rfloor$ for the largest integer not exceeding N^*.

Proposition 7 (Endogenous entry in the frictionless benchmark)**.** *Suppose $k = 0$, $\lambda = 1$, $\ell > s$,* *and* $0 < \kappa < A$. *The per-firm equilibrium profit schedule*

$$\Pi^*(N) = \begin{cases} A/N & \text{if } N \leq N^*, \\ A/N - \Delta & \text{if } N > N^*, \end{cases}$$

is strictly decreasing on $\mathbb{N}$. *The unique pure-strategy free-entry equilibrium has* $N^{FE} = \max\{N \in \mathbb{N} : \Pi^*(N) \geq \kappa\}$ *firms. Three regimes arise depending on the entry cost:*

(i) Low entry cost *(*$\kappa \leq A/(m+1) - \Delta$*):* $N^{FE} = \lfloor A/(\kappa + \Delta) \rfloor \geq m + 1$. *Every firm fully automates.*

(ii) Intermediate entry cost *(*$A/(m+1) - \Delta < \kappa < A/m$*):* $N^{FE} = m$. *No firm automates, yet each earns strictly positive profit* $A/m > \kappa$. *The* threat *of automation deters further entry: one additional firm would trigger full automation, dropping per-firm profit to* $A/(m+1) - \Delta < \kappa$.

(iii) High entry cost *(*$\kappa \geq A/m$*):* $N^{FE} = \lfloor A/\kappa \rfloor \leq m$. *No firm automates. Entry costs, not the automation externality, limit the number of firms.*

When case (i) arises, the Prisoner's Dilemma of Corollary 1 materializes under free entry: all firms automate, demand contracts, and every firm would be better off had none automated. Case (iii) is the standard free-entry outcome: entry costs are high enough that the market never approaches the automation threshold, and the externality is irrelevant. Case (ii) is the most distinctive: the *threat* of automation functions as an endogenous entry barrier, sustaining positive profits without any automation actually occurring, at the cost of sustaining market power.

The convex-cost case is less stark but more robust. With $k > 0$, the automation rate $\alpha^{NE}(N) = (s - \ell/N)/k$ varies continuously in N, so the profit schedule no longer jumps at N^* and the entry-deterrence mechanism of Proposition 7(ii) does not arise.

Proposition 8 (Endogenous entry with convex costs)**.** *Suppose* $k > 0$, $\ell > s$, *and* $\Pi_0(1) > \kappa$ *(the market is viable). A free-entry equilibrium* N^{FE} *satisfying* (11) *exists. If* $N^{FE} > N^*$, *then* $\alpha^{NE}(N^{FE}) > \alpha^{CO}$*: over-automation persists under free entry.*

Generically, N^{FE} exceeds N^* whenever zero-automation profits at the threshold, $\Pi_0(N^*)$, exceed κ.[3] Free entry then pins down the number of firms but does not alter the strategic incentives within the automation subgame: each firm still bears only a fraction $1/N^{FE}$ of the demand loss,

---

[3] In a numerical grid over $c/w \in \{0.1, \ldots, 0.5\}$, $\lambda \in \{0.3, \ldots, 1\}$, $\eta \in \{0, \ldots, 0.3\}$, $k \in \{0.5, 1, 2\}$, and $\kappa \in \{0.1, \ldots, 5\}$, N^{FE} exceeds N^* in over 94% of parameterizations satisfying the proposition's conditions; the exceptions arise only when the entry cost is high enough that the market barely supports more than N^* firms.

and the over-automation wedge $\ell(1-1/N^{FE})/k$ persists. If $N^{FE} \leq N^*$, no firm automates and the outcome is efficient, but only because the market is too concentrated for the private automation incentive to activate.

Taken together, the two propositions deliver a common lesson: free entry reshapes the over-automation problem but does not resolve it. If anything, the standard tendency toward excess entry (Mankiw and Whinston, 1986) widens the wedge by fragmenting the market further.

The entry margin also reveals an unintended side effect of UBI (Section 4.2). By raising autonomous demand A, UBI increases per-firm profit $\Pi_0 = A/N + (\lambda - 1)wL$ at any given N, attracting additional entrants until the zero-profit condition (11) binds at a larger N^{FE}. Since the over-automation wedge $\ell(1-1/N)/k$ is increasing in N, a policy designed to cushion displacement can paradoxically widen the very externality that causes it.

5.3 Endogenous Wages

A central insight of Acemoglu and Restrepo (2018) is that endogenous wage adjustment can stabilize the automation path: as firms automate, displaced workers increase labor supply, pushing wages down; lower wages narrow the cost saving from automation and discourage further displacement. This self-correcting feedback is a natural candidate for resolving the demand externality identified above. We show that it raises the threshold at which the externality activates but cannot close the wedge once it does.

In Acemoglu and Restrepo (2018), wages are determined by labor-market clearing in a full general equilibrium; we adopt a reduced-form representation that captures the key qualitative feature of their mechanism. Let the wage depend on the aggregate automation rate: $w(\bar{\alpha})$ with $w(\bar{\alpha}) > c$ and $w'(\bar{\alpha}) \leq 0$; firms take the prevailing wage as given when choosing α_i. This specification requires only that wages fall when aggregate labor-market slack increases, a property shared by efficiency-wage models, where the no-shirking wage declines with unemployment (Shapiro and Stiglitz, 1984), and by the empirical wage curve, where Blanchflower and Oswald (1995) document a robust negative relationship between wages and unemployment across more than a dozen countries. In our setting, automation displaces workers into the labor pool, generating exactly this type of slack.

Both the cost saving $s(w) = w - c$ and the demand-loss parameter $\ell(w) = \lambda(1-\eta)w$ are increasing in the wage, so falling wages affect both sides of the automation margin: they shrink the private incentive to automate (the self-correcting channel) and reduce the demand loss per automated task. The equilibrium automation rate is therefore a fixed point in which automation, wages, and the externality are jointly determined. Despite this richer feedback, the threshold

$N^* = \ell/s = \lambda(1-\eta)w/(w-c)$ rises as wages fall, because the cost saving $s = w - c$ contracts faster than the demand loss $\ell = \lambda(1-\eta)w$. The structural source of the distortion, however, is unaffected.

Proposition 9 (Robustness to wage adjustment)**.** *Let* $k > 0$*, let* $w : [0, 1] \to (c, \infty)$ *be differentiable with* $w'(\bar{\alpha}) \leq 0$*, and suppose firms are wage-takers.*

(i) At any symmetric equilibrium with $N > N^*(w(\bar{\alpha}))$*, the Nash automation rate exceeds the cooperative optimum:* $\alpha^{NE} > \alpha^{CO}$*.*

(ii) Endogenous wage adjustment raises the threshold: $N^*(w(\bar{\alpha})) \geq N^*(w(0))$ *for all* $\bar{\alpha} \in [0, 1]$*, with strict inequality whenever* $w(\bar{\alpha}) < w(0)$*.*

Competitive pricing allocates revenue as $\text{Rev}_i = D/N$ at any wage level, so each firm bears only a fraction of the demand destruction its automation causes regardless of whether w is high or low. Wage adjustment changes the *magnitude* of ℓ but not the *fraction* each firm internalizes; that fraction is a property of market structure, not of factor prices.

The strongest version of the self-correcting argument is that wages could fall far enough to shut the externality down entirely. As $w \to c$, the cost saving $s \to 0$ while $N^* \to \infty$: eventually N^* exceeds N and no firm finds automation privately worthwhile. But this is a Pyrrhic resolution. When wages are driven to near the AI cost, workers who *retain* their jobs earn little more than the machines that would replace them, and aggregate purchasing power collapses through wage depression rather than displacement. The externality vanishes not because the demand problem has been solved but because there is so little income left per worker that the wedge between private and social incentives becomes negligible: a labor market that "self-corrects" only by impoverishing its workforce has transmuted displacement into depressed living standards. More generally, wage flexibility changes *when* the externality bites, not *whether* it exists.

The analysis above sets $\mu = 0$, so the planner cares only about firm profits. A planner who also values worker welfare ($\mu > 0$) would find wage depression no more acceptable than displacement, demanding a larger correction, yet wage adjustment provides the same compression of ℓ. Endogenous wages therefore close a smaller share of the gap, making them even less adequate as a corrective mechanism. Corollary 4 in the appendix confirms this: the over-automation result extends to any $\mu \in [0, 1)$ under endogenous wages.

5.4 Capital Income Recycling

Section 4.1 showed that raising η, the fraction of displaced income recovered by workers, shrinks the demand-loss parameter ℓ and narrows the over-automation wedge. A natural counterpart on the capital side is that owners spend their profits: if their consumption offsets the spending lost through displacement, the demand externality might disappear. We show that recycling narrows the wedge but cannot close it under empirically plausible parameters.

Suppose capital owners consume a fraction $\hat{\eta} \in [0, 1)$ of their capital income in the sector. Total sector profit is $\Pi = D - NL(w - s\bar{\alpha})$. Adding capital consumption $\hat{\eta}\Pi$ to aggregate demand and solving for D yields

$$D = \frac{A + (\lambda - \hat{\eta})wLN}{1 - \hat{\eta}} - \frac{\ell_{\hat{\eta}}LN}{1 - \hat{\eta}}\,\bar{\alpha}, \tag{12}$$

where

$$\ell_{\hat{\eta}} = \ell - \hat{\eta}s$$

is the effective demand-loss parameter: each automated task loses ℓ in worker spending, but owners recycle $\hat{\eta}$ of the per-task saving s back into demand. When $\hat{\eta} = 0$, $\ell_{\hat{\eta}} = \ell$ and eq. (12) reduces to eq. (2).

Competitive pricing still gives $\text{Rev}_i = D/N$. The first-order condition becomes $L(s - \ell_{\hat{\eta}}/[N(1 - \hat{\eta})])$, giving a modified threshold

$$N_{\hat{\eta}} := \frac{\ell_{\hat{\eta}}}{s(1 - \hat{\eta})}.$$

Proposition 10 (Capital income recycling)**.** *Suppose $k = 0$ and $\ell_{\hat{\eta}} > 0$. When there is capital income recycling at rate $\hat{\eta}$,*

(i) full automation is dominant if and only if $N > N_{\hat{\eta}}$.

(ii) The externality vanishes ($\ell_{\hat{\eta}} \leq 0$) only when $\hat{\eta} \geq \ell/s$.

Part (ii) requires $\hat{\eta} \geq \ell/s = \lambda(1 - \eta)w/(w - c)$: owners must recycle enough of each task's cost saving to replace the demand that displaced workers would have generated. When $\ell > s$, the required rate exceeds one, so recycling is impotent precisely where the externality is most harmful, since $\ell > s$ implies $\alpha^{CO} = 0$ and firms automate when the planner would prefer none. When $\ell < s$, elimination is feasible in principle, but the planner already prefers positive automation (Proposition 1) and the wedge is quantitatively smaller.

The frictionless case gives the sharpest result, but the structure carries over when frictions are positive. The proof of Proposition 10 extends the result to $k > 0$. The Nash equilibrium generalizes to $\alpha^{NE} = (s - \ell/\hat{N})/k$, where $\hat{N} = N(1 - \hat{\eta}) + \hat{\eta}$ is an effective market size that interpolates between N (no recycling) and 1 (full recycling), making each firm behave as though it faced fewer competitors. The cooperative optimum, however, is unchanged: the $1/(1 - \hat{\eta})$ multiplier scales total profit without shifting the optimizer.

The upshot parallels Section 4.1: recycling raises the fraction of demand loss each firm internalizes from $1/N$ to $1/\hat{N}$, but cannot push it to one. Addressing how income is spent narrows the wedge but does not close it, because the underlying dilution across firms persists.

5.5 Imperfect Product-Market Competition and Task Complementarity

The baseline assumes competitive pricing and perfect substitution across tasks. Without providing a formal treatment, we argue that richer product-market interaction and task complementarity would complicate the analysis but should not eliminate the demand externality.

Second-stage price or quantity competition. If firms competed on quantities (Cournot) or prices (Bertrand) after choosing automation rates, two new forces would arise. The first is a demand-allocation effect: strategic competition changes how a given level of expenditure is divided among firms. The second is a market-share motive: a firm that automates more than its rivals can undercut on price or expand output, capturing a larger share of expenditure. These forces have different implications for the size of the over-automation wedge, but neither eliminates the demand externality itself, because neither changes the fact that automation reduces the *level* of aggregate expenditure D through worker displacement.

The demand-allocation effect is straightforward. Whether revenue is split by competitive pricing, Cournot market shares, or Bertrand undercutting, each firm would still bear only a fraction of the reduction in D that its automation causes. With differentiated products, the uninternalized demand loss scales with the share of the market the firm does *not* capture; over-automation therefore persists whenever a firm lacks full market power.

The market-share motive is more subtle. At the symmetric equilibrium, market-share gains cancel across firms, echoing the Red Queen structure of Section 5.1. Under Cournot, however, the firm that expands its market share also absorbs a larger fraction of the demand loss it causes, partially offsetting the motive; the net effect on the wedge likely depends on market structure details.

CES task aggregation. The task-based framework of Acemoglu and Restrepo (2018) aggregates tasks via a CES production function; our baseline takes the perfect-substitutes limit for tractability. Under the general CES aggregator with an elasticity of substitution greater than one (tasks are substitutes but not perfect), automating the marginal task yields diminishing output gains, creating an interior optimum even without the convex integration cost k. Because this changes the shape of the production function but not the demand side, the demand externality still operates at the automation margin and the over-automation wedge persists, though it is smaller because diminishing returns already restrain automation from the supply side.

When tasks are complements (elasticity less than one), the restraint is stronger: the production function itself limits automation, since eliminating human tasks reduces output. Even so, the demand externality remains positive as long as displaced workers lose income.

In each case, richer modeling assumptions would change the magnitude of over-automation but not its source: the wedge persists as long as firms do not fully internalize the demand loss from displacement. Characterizing the exact interaction between strategic pricing, market-share motives, and the demand externality is a promising direction for future work; the baseline's simplifying assumptions isolate the core mechanism in its starkest form.

6 Discussion

This paper develops a simple model with a simple but stark insight. Even as AI-driven layoffs sweep across industries, and even as every firm recognizes that vanishing paychecks mean vanishing customers, not one of them will stop. Each firm reaps the full savings of replacing its own workers yet bears only a sliver of the demand it destroys; the rest lands on rivals. No firm can afford to be the one that holds back. This is the trap: an automation arms race that only intensifies as AI improves, that leaves workers and firm owners alike worse off, and that no market force can break. We close by discussing implications for empirics and policy, and then the scope and limitations of the analysis.

Empirical implications. Anthropic CEO Dario Amodei has warned that AI-driven displacement will be "unusually painful," "much broader" and "much faster" than previous technological shocks (CNBC, 2026a). If that assessment proves correct and income replacement remains incomplete, the model points, perhaps counterintuitively, to where the problem is most severe: not dominant technology firms but fragmented industries deploying the most capable AI (Propositions 1 and 6). The distinguishing empirical signature would be profit erosion. Standard competitive models predict that cost-reducing technology raises profits; profit erosion that coincides with mass layoffs would be difficult to rationalize without the externality (Proposition 2). That said, this signature requires displacement at a scale and speed beyond what has materialized so far. If reabsorption keeps pace with automation, the externality may remain too small to detect, and the paper's contribution is identifying a structural vulnerability rather than diagnosing an active crisis. Three settings where AI-driven displacement is already under way offer concrete starting points: customer support, where thousands of firms are simultaneously replacing agents with agentic AI (CNBC, 2025c); software services, where tools that enable one engineer to replace a multi-person team (CNBC, 2025a) create measurable shifts in headcount-to-output ratios; and back-office operations across competing financial institutions, where regulatory reporting makes both adoption rates and revenue outcomes unusually transparent.

More broadly, the literature has shown that AI delivers large productivity gains (Brynjolfsson et al., 2025b), that competitive pressure accelerates adoption (Li et al., 2025), and that effective human oversight grows harder as AI improves (Bastani and Cachon, 2025). Each of these findings documents a strong firm-level incentive to automate. Our model shows what happens when every firm acts on that incentive at once: the returns any one firm observes do not account for the demand that all firms collectively destroy, so private returns to AI systematically overstate the returns to the economy as a whole.

Policy implications. Much of the policy debate around AI-driven displacement focuses on how to respond after the fact, through retraining, income support, or regulation. Our results reframe the question: do competitive incentives drive firms to automate beyond what is collectively optimal? Even a planner who places zero weight on worker welfare would reduce the automation rate below the equilibrium level (Proposition 2). The problem is not that firms profit at workers' expense; it is that over-automation harms both groups, making correction a matter of eliminating waste rather than redistributing gains. Universal basic income, perhaps the most widely discussed response, raises living standards but does not change a single firm's incentive to automate (Section 4.2). Collective bargaining faces the same wall: because automation is a dominant strategy, no voluntary agreement among firms to restrain layoffs is self-enforcing (Section 4.5). By Tinbergen's principle, a distinct market failure requires a distinct instrument; only a Pigouvian automation tax supplies it (Table 1). No amount of retraining, income support, or bargaining will slow the arms race; only a tax on automation itself changes the calculus that drives it.

One practical consideration bears on implementation: the model is a closed-sector game, and a unilateral automation tax could push adoption offshore, strengthening the case for multilateral coordination or border-adjustment mechanisms analogous to those used in carbon policy.

Scope, limitations, and future directions. The model is deliberately simple: one sector, one period, symmetric firms. Each of these choices is conservative, meaning the real problem is likely worse than what we show.

A single sector understates the externality. In a multi-sector economy, layoffs in one sector reduce spending on every sector's output, creating reinforcing demand spirals. Platform ecosystems make the point concrete: when a platform automates seller support, gig logistics, or content moderation, the lost spending cascades across an entire ecosystem of complementors.

The static setting misses two dynamic forces that pull in opposite directions. AI investments are largely irreversible, and Proposition 7 shows that even the *threat* of automation can reshape market structure before any displacement occurs, strengthening the case for early policy action. Working the other way, the income-replacement rate η rises over time as displaced workers retrain and new occupations emerge (Acemoglu and Restrepo, 2019), so the optimal tax should shrink as the economy adjusts (Section 4.6).

Symmetry rules out heterogeneity across firms and workers, and endogenizing AI development could compound the problem: firms racing to automate may invest disproportionately in labor-replacing AI rather than labor-augmenting AI (Acemoglu and Restrepo, 2018), feeding the very arms race the model identifies.

Each of these extensions points in the same direction: toward a larger problem, not a smaller one. Nonetheless, pursuing them, along with the empirical tests outlined above, are promising directions for future work.

References

Daron Acemoglu. The simple macroeconomics of AI. *Economic Policy*, 40(121):13–58, 2025.

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.

Daron Acemoglu and Pascual Restrepo. The wrong kind of AI? Artificial intelligence and the future of labour demand. *Cambridge Journal of Regions, Economy and Society*, 13(1):25–35, 2020.

Daron Acemoglu and Pascual Restrepo. Automation and rent dissipation: Implications for wages, inequality, and productivity. *Quarterly Journal of Economics*, forthcoming.

Imanol Arrieta-Ibarra, Leonard Goff, Diego Jiménez-Hernández, Jaron Lanier, and E. Glen Weyl. Should we treat data as labor? Moving beyond "free". *AEA Papers and Proceedings*, 108:38–42, 2018.

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

David H. Autor, Caroline Chin, Anna Salomons, and Bryan Seegmiller. New frontiers: The origins and content of new work, 1940–2018. *Quarterly Journal of Economics*, 139(3):1399–1465, 2024.

Martino Banchio and Giacomo Mantegazza. Artificial intelligence and spontaneous collusion. Working paper, Bocconi University and USC Marshall School of Business, 2022.

Hamsa Bastani and Gérard P. Cachon. The human-AI contracting paradox. Working paper, The Wharton School, University of Pennsylvania, December 2025. SSRN 5962739.

Seth G. Benzell, Laurence J. Kotlikoff, Guillermo LaGarda, and Jeffrey D. Sachs. Robots are Us: Some economics of human replacement. Working Paper 20941, National Bureau of Economic Research, February 2015.

Martin Beraja and Nathan Zorzi. Inefficient automation. *Review of Economic Studies*, 92(1):69–96, 2025.

David G. Blanchflower and Andrew J. Oswald. An introduction to the wage curve. *Journal of Economic Perspectives*, 9(3):153–167, 1995.

Erik Brynjolfsson and Andrew McAfee. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton, New York, 2014.

Erik Brynjolfsson, Bharat Chandar, and Ruyu Chen. Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. Working paper, Stanford Institute for Economic Policy Research, August 2025a.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *Quarterly Journal of Economics*, 140 (2):889–942, 2025b.

CNBC. Goldman Sachs is piloting its first autonomous coder in major AI milestone for Wall Street. `https://www.cnbc.com/2025/07/11/goldman-sachs-autonomous-coder-pilot-marks-major-ai-milestone.html`, 2025a. Accessed February 28, 2026.

CNBC. AI job cuts: Amazon, Microsoft, and more cite AI for 2025 layoffs. `https://www.cnbc.com/2025/12/21/ai-job-cuts-amazon-microsoft-and-more-cite-ai-for-2025-layoffs.html`, 2025b. Accessed February 28, 2026.

CNBC. Salesforce CEO confirms 4,000 layoffs “because I need less heads” with AI. `https://www.cnbc.com/2025/09/02/salesforce-ceo-confirms-4000-layoffs-because-i-need-less-heads-with-ai.html`, 2025c. Accessed February 28, 2026.

CNBC. Anthropic CEO Dario Amodei warns AI may cause ‘unusually painful’ disruption to jobs. `https://www.cnbc.com/2026/01/27/dario-amodei-warns-ai-cause-unusually-painful-disruption-jobs.html`, 2026a. Accessed March 20, 2026.

CNBC. Block laying off about 4,000 employees, nearly half of its workforce. `https://www.cnbc.com/2026/02/26/block-laying-off-about-4000-employees-nearly-half-of-its-workforce.html`, 2026b. Accessed February 28, 2026.

R. H. Coase. The problem of social cost. *Journal of Law and Economics*, 3:1–44, 1960.

Russell Cooper and Andrew John. Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economics*, 103(3):441–463, 1988.

Arnaud Costinot and Iván Werning. Robots, trade, and Luddism: A sufficient statistic approach to optimal technology regulation. *Review of Economic Studies*, 90(5):2261–2291, 2023.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: Labor market impact potential of large language models. *Science*, 384:1306–1308, 2024.

João Guerreiro, Sergio Rebelo, and Pedro Teles. Should robots be taxed? *Review of Economic Studies*, 89(1):279–311, 2022.

Daniel S. Hamermesh and Gerard A. Pfann. Adjustment costs in factor demand. *Journal of Economic Literature*, 34 (3):1264–1292, 1996.

Infosys. Infosys and Cognition announce strategic collaboration to accelerate the AI value journey for global enterprises. Press Release, January 7, 2026. `https://www.prnewswire.com/news-releases/infosys-and-cognition-announce-strategic-collaboration-to-accelerate-the-ai-value-journey-for-global-enterprise` `html`, accessed March 20, 2026.

Louis S. Jacobson, Robert J. LaLonde, and Daniel G. Sullivan. Earnings losses of displaced workers. *American Economic Review*, 83(4):685–709, 1993.

Nicholas Kaldor. Alternative theories of distribution. *Review of Economic Studies*, 23(2):83–100, 1956.

Jussi Keppo, Yuze Li, Gerry Tsoukalas, and Nuo Yuan. On the fragility of AI agent collusion. Working paper, Boston University and National University of Singapore, January 2026. SSRN 5386338.

John Maynard Keynes. Economic possibilities for our grandchildren. In *Essays in Persuasion*, pages 358–373. Macmillan, London, 1930.

Anton Korinek and Joseph E. Stiglitz. Artificial intelligence and its implications for income distribution and unemployment. In Ajay Agrawal, Joshua Gans, and Avi Goldfarb, editors, *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, Chicago, 2019.

Wassily Leontief. The distribution of work and income. *Scientific American*, 247(3):188–204, 1982.

Boshuo Li, Ni Huang, and Wei Shi. Forced to change? Media exposure of labor issues and firm artificial intelligence investment. *Information Systems Research*, 2025. doi: 10.1287/isre.2022.0402. Articles in Advance.

Robert E. Lucas. Adjustment costs and the theory of supply. *Journal of Political Economy*, 75(4):321–334, 1967.

N. Gregory Mankiw and Michael D. Whinston. Free entry and social inefficiency. *RAND Journal of Economics*, 17 (1):48–58, 1986.

Atif Mian, Ludwig Straub, and Amir Sufi. Indebted demand. *Quarterly Journal of Economics*, 136(4):2243–2307, 2021.

Kevin M. Murphy, Andrei Shleifer, and Robert W. Vishny. Industrialization and the big push. *Journal of Political Economy*, 97(5):1003–1026, 1989.

Arthur Cecil Pigou. *The Economics of Welfare*. Macmillan, London, 1920.

Eric A. Posner and E. Glen Weyl. *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton University Press, Princeton, 2018.

David Ricardo. *On the Principles of Political Economy and Taxation*. John Murray, London, 3rd edition, 1821.

Paul N. Rosenstein-Rodan. Problems of industrialisation of Eastern and South-Eastern Europe. *Economic Journal*, 53(210/211):202–211, 1943.

Alap Shah. The 2028 Global Intelligence Crisis. Citrini Research, February 2026. URL `https://www.citriniresearch.com/p/2028gic`.

Carl Shapiro and Joseph E. Stiglitz. Equilibrium unemployment as a worker discipline device. *American Economic Review*, 74(3):433–444, 1984.

Martin L. Weitzman. The simple macroeconomics of profit sharing. *American Economic Review*, 75(5):937–953, 1985.

Joseph Zeira. Workers, machines, and economic growth. *Quarterly Journal of Economics*, 113(4):1091–1117, 1998.

A Proofs

Proof of Proposition 1.

Preamble and part (i): Nash equilibrium. From eq. (6), we have

$$\pi_i = \Pi_0 + L\left[\alpha_i\left(s - \frac{\ell}{N}\right) - \frac{k}{2}\alpha_i^2 - \frac{\ell}{N}\sum_{j\neq i}\alpha_j\right]$$

Differentiating with respect to α_i (only the terms involving α_i contribute, since the rivals' sum $\sum_{j\neq i}\alpha_j$ is a constant from firm i's perspective):

$$\frac{\partial \pi_i}{\partial \alpha_i} = L\left[\left(s - \frac{\ell}{N}\right) - k\alpha_i\right]. \tag{13}$$

The second derivative is

$$\frac{\partial^2 \pi_i}{\partial \alpha_i^2} = -kL < 0,$$

so π_i is strictly concave in α_i.

Setting the first-order condition to zero:

$$L\left(s - \frac{\ell}{N} - k\alpha_i\right) = 0 \qquad \Longrightarrow \qquad \alpha_i^* = \frac{s - \ell/N}{k}.$$

Now, α_i is restricted to the range $[0, 1]$ (since it is the fraction of work automated).

If $\alpha_i^* > 1$, then since $\frac{\partial^2 \pi_i}{\partial \alpha_i^2} < 0$, $\frac{\partial \pi_i}{\partial \alpha_i} > 0$ for $\alpha_i < \alpha_i^*$, so in this case, the maximum value of $\pi_i(\alpha_i)$ occurs at $\alpha_i = 1$.

Similarly, if $\alpha_i^* < 0$, then $\frac{\partial \pi_i}{\partial \alpha_i} < 0$ for $\alpha_i > \alpha_i^*$, so in this case, the maximum value of $\pi_i(\alpha_i)$ occurs at $\alpha_i = 0$. Note that $\alpha_i^* < 0$ if and only if $s < \ell/N$, i.e., $N < \ell/s = N^*$; this establishes the preamble claim that no firm automates when $N \leq N^*$.

Note that α_i^* does not depend on any rival's choice α_j ($j \neq i$): the rivals' automation levels enter only through the additive term $-(\ell/N)\sum_{j\neq i}\alpha_j$, which does not affect the first-order condition. Hence firm i's optimal strategy is independent of α_{-i} (a dominant strategy), and is to set

$$\alpha_i = \frac{s - \ell/N}{k},$$

provided the right-hand side lies in $[0, 1]$; otherwise $\alpha_i = 0$ or $\alpha_i = 1$ at the respective boundary. Since every firm solves the same problem, the unique symmetric Nash equilibrium has $\alpha_i = \alpha^{NE}$ for all i.

Part (ii): Cooperative optimum. Summing the per-firm profit (5) over all N firms, the sector's total profit is

$$\begin{aligned}\pi_{tot} &\coloneqq \sum_{i=1}^{N} \pi_i \\ &= \sum_{i=1}^{N} \left[\Pi_0 + L\left(s\alpha_i - \ell\bar{\alpha} - \frac{k}{2}\alpha_i^2\right)\right] \\ &= N\Pi_0 + Ls\sum_i \alpha_i - L\ell N\bar{\alpha} - \frac{kL}{2}\sum_{i=1}^{N}\alpha_i^2.\end{aligned}$$

Using $\sum_i \alpha_i = N\bar{\alpha}$, this simplifies to

$$\pi_{tot} = N\Pi_0 + NL\bar{\alpha}(s-\ell) - \frac{kL}{2}\sum_{i=1}^{N}\alpha_i^2.$$

The first two terms depend on $\boldsymbol{\alpha}$ only through $\bar{\alpha}$; the last term, $-\frac{kL}{2}\sum_i \alpha_i^2$, is maximized (least negative) when the α_i are as equal as possible, by convexity of x^2. Formally, for any fixed $\bar{\alpha}$, $\sum_i \alpha_i^2 \geq N\bar{\alpha}^2$ with equality if and only if $\alpha_i = \bar{\alpha}$ for all i (by the QM–AM inequality). Hence the optimum is symmetric: $\alpha_i = \alpha$ for all i, and $\bar{\alpha} = \alpha$. Substituting:

$$\pi_{tot} = N\Pi_0 + NL\alpha(s-\ell) - \frac{kNL\alpha^2}{2}. \tag{14}$$

This is a concave quadratic in α:

$$\frac{\partial \pi_{tot}}{\partial \alpha} = NL(s-\ell) - kNL\alpha, \qquad \frac{\partial^2 \pi_{tot}}{\partial \alpha^2} = -kNL < 0.$$

Setting the first-order condition to zero:

$$NL(s-\ell) - kNL\alpha = 0 \quad \Rightarrow \quad \alpha^* = \frac{s-\ell}{k}.$$

Since π_{tot} is strictly concave, this is the unique global maximum on $\mathbb{R}$. Restricting to $\alpha \in [0,1]$: if $(s-\ell)/k > 1$, strict concavity implies $\partial\pi_{tot}/\partial\alpha > 0$ on $[0,1]$, so the constrained maximum is $\alpha = 1$; if $(s-\ell)/k < 0$, then $\partial\pi_{tot}/\partial\alpha < 0$ on $[0,1]$, so the constrained maximum is $\alpha = 0$. Thus

$$\alpha^{CO} = \frac{s-\ell}{k},$$

with boundary values 0 or 1 when the expression falls outside $[0,1]$.

Part (iii): Over-automation wedge. Suppose $N > N^*$, so $\alpha^{NE} = (s - \ell/N)/k > 0$ by (i).

Case $s > \ell$. Both $(s-\ell/N)/k > 0$ and $(s-\ell)/k > 0$, so neither rate is clamped at zero. Subtracting:

$$\alpha^{NE} - \alpha^{CO} = \frac{s-\ell/N}{k} - \frac{s-\ell}{k} = \frac{\ell - \ell/N}{k} = \frac{\ell\,(1-1/N)}{k}.$$

This is strictly positive for $N \geq 2$ because $\ell > 0$ and $1 - 1/N > 0$.

For the comparative statics, write the wedge as $W = \ell(1-1/N)/k$:

$$\begin{aligned}
\frac{\partial W}{\partial N} &= \frac{\ell}{kN^2} > 0 && \text{(increasing in } N\text{)},\\
\frac{\partial W}{\partial \ell} &= \frac{1-1/N}{k} > 0 && \text{(increasing in } \ell\text{)},\\
\frac{\partial W}{\partial k} &= -\frac{\ell(1-1/N)}{k^2} < 0 && \text{(decreasing in } k\text{)}.
\end{aligned}$$

Part (iv): Boundary cases.

When $s \leq \ell$, then $(s-\ell)/k \leq 0$, so $\alpha^{CO} = 0$. So the wedge is $\alpha^{NE} - \alpha^{CO} = \alpha^{NE}$. Since we are in the setting where $N > N^*$, we have $(s-\ell/N)/k > 0$. It is still possible, however, that $(s-\ell/N)/k > 1$.

If $s < k + \ell/N$, then $\alpha^{NE} = (s-\ell/N)/k$, and this is the wedge. If $s > k + \ell/N$, then α^{NE} is a corner case, $\alpha^{NE} = 1$, and this is the wedge. □

Lemma 1 (Boundary cases)**.** *Write $\delta(\mu) := \frac{\mu\,\ell}{\lambda(1-\mu)}$ for the additional planner correction beyond α^{CO}. Then, for $k > 0$*

(i) α^{NE} is interior if and only if $\ell/N < s < k+\ell/N$; Note that $N > N^*$ is equivalent to $s > \ell/N$.

(ii) α^{CO} is interior if and only if $\ell < s < k + \ell$;

(iii) $\alpha^{SP}(\mu)$ is interior if and only if $\ell + \delta(\mu) < s < k + \ell + \delta(\mu)$.

Each interval has width k. Because $\ell/N < \ell < \ell+\delta(\mu)$ for $\mu > 0$, the three windows are progressively shifted to the right: interior automation arises at the lowest cost savings under Nash behavior, at intermediate savings under cooperation, and only at the highest savings under the social planner.

***Proof of Lemma 1*.** Each claim follows by checking when the raw formula lies strictly in $(0,1)$.

(i) Nash equilibrium. From Proposition 1(i), $\alpha^{NE} = (s-\ell/N)/k$. The raw formula $(s-\ell/N)/k$ is strictly positive iff $s > \ell/N$ (i.e., when $N > N^*$), and strictly less than one iff $s < k + \ell/N$.

(ii) Cooperative optimum. From Proposition 1(ii), $\alpha^{CO} = (s-\ell)/k$. The raw expression $(s-\ell)/k$ is strictly positive iff $s > \ell$, and strictly less than one iff $s < k + \ell$.

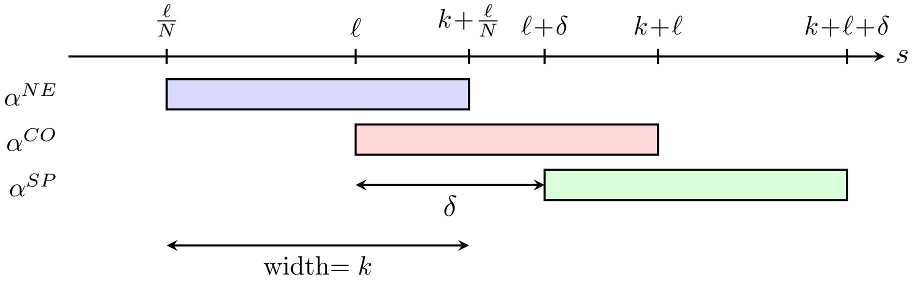

Figure 3: Interior regions for α^{NE}, α^{CO}, and $\alpha^{SP}(\mu)$ as functions of the cost saving s, where $\delta = \mu\ell/[\lambda(1-\mu)]$. Each shaded bar spans the values of s for which the corresponding automation rate is strictly interior. All three intervals have the same width k; they are shifted rightward by the progressively larger demand-loss terms that each objective internalizes.

(iii) Social planner. From Proposition 2(i), $\alpha^{SP}(\mu) = (s-\ell)/k - \mu\ell/[\lambda(1-\mu)k]$. Collecting terms, $\alpha^{SP}(\mu) = \big(s-\ell-\delta(\mu)\big)/k$, which is strictly positive iff $s > \ell + \delta(\mu)$ and strictly less than one iff $s < k + \ell + \delta(\mu)$.

Finally, $\ell/N < \ell$ holds because $N > N^* \geq 1$ (indeed $N \geq 2$ in any strategic setting), and $\ell < \ell + \delta(\mu)$ for $\mu > 0$ since $\delta(\mu) > 0$. □

Proof of Corollary 1 *(Frictionless limit, corollary of Proposition 1).*

(i) With $k = 0$, firm i's profit from (6) becomes

$$\pi_i = \Pi_0 + L\left[\alpha_i\left(s - \frac{\ell}{N}\right) - \frac{\ell}{N}\sum_{j\neq i}\alpha_j\right].$$

This is affine in α_i, with slope $L(s-\ell/N)$. Crucially, the slope does not depend on rivals' choices α_{-i} — rival automation affects the level of firm i's profit (through the last term) but not the marginal return to α_i. The optimal α_i therefore does not depend on α_{-i}, making it a dominant strategy.

When $N > N^*$ (equivalently $s > \ell/N$), the slope is strictly positive. Since π_i is linear and increasing in α_i on $[0,1]$, the unique optimum is the upper boundary $\alpha_i^* = 1$. This holds for every firm simultaneously, so $\alpha_i = 1$ for all i is the unique Nash equilibrium.

(ii) Assume additionally that the cost saving is less than the demand loss: $s < \ell$. With $k = 0$, aggregate profit from (5) is linear in $\bar{\alpha}$: $\sum_i \pi_i = N\Pi_0 + NL(s-\ell)\bar{\alpha}$. Since $s - \ell < 0$, this is strictly decreasing in $\bar{\alpha}$, so the profit-maximizing cooperative outcome is $\bar{\alpha} = 0$ (no automation), yielding per-firm profit Π_0.

At the Nash equilibrium ($\bar{\alpha} = 1$), per-firm profit is $\Pi_0 + L(s - \ell)$. Since $s - \ell < 0$, we have $\Pi_0 + L(s - \ell) < \Pi_0$: every firm earns strictly less than under cooperation. The per-firm profit loss is $\Pi_0 - [\Pi_0 + L(s - \ell)] = L(\ell - s)$. Across N firms, the total deadweight loss is $NL(\ell - s)$. For the demand loss, from eq. (2):

$$D = A + \lambda wLN - \ell LN\,\bar{\alpha}.$$

At $\bar{\alpha} = 0$: $D^{CO} = A + \lambda wLN$. At $\bar{\alpha} = 1$: $D^{NE} = A + \lambda wLN - \ell LN$. Hence $D^{CO} - D^{NE} = \ell LN$.

$\square$

Proof of Proposition 2.

(i) The μ-planner chooses $\bar{\alpha}$ to maximize $S(\mu) = \mu\,\mathcal{W} + (1 - \mu)\,\mathcal{K}$ from (8). We compute each derivative in turn. Recall that $\mathcal{W} = wLN[1 - (1 - \eta)\bar{\alpha}]$, so

$$\frac{d\mathcal{W}}{d\bar{\alpha}} = -wNL(1 - \eta). \tag{15}$$

Since $\ell = \lambda(1 - \eta)w$, this equals $-\ell\,NL/\lambda$. Recall that $\mathcal{K} = N[\Pi_0 + L((s - \ell)\bar{\alpha} - \frac{k}{2}\bar{\alpha}^2)]$ (e.g. eq. (14)), so

$$\frac{d\mathcal{K}}{d\bar{\alpha}} = NL\big[(s - \ell) - k\bar{\alpha}\big]. \tag{16}$$

Combining:

$$\frac{dS}{d\bar{\alpha}} = \mu\,\frac{d\mathcal{W}}{d\bar{\alpha}} + (1 - \mu)\,\frac{d\mathcal{K}}{d\bar{\alpha}} = -\frac{\mu\,\ell\,NL}{\lambda} + (1 - \mu)\,NL\big[(s - \ell) - k\bar{\alpha}\big].$$

Setting this to zero and dividing by $NL > 0$:

$$(1 - \mu)\big[(s - \ell) - k\bar{\alpha}\big] = \frac{\mu\,\ell}{\lambda}.$$

Dividing both sides by $(1 - \mu) > 0$ and isolating $k\bar{\alpha}$:

$$(s - \ell) - k\bar{\alpha} = \frac{\mu\,\ell}{\lambda(1 - \mu)}, \qquad \Longrightarrow \qquad k\bar{\alpha} = (s - \ell) - \frac{\mu\,\ell}{\lambda(1 - \mu)}.$$

Dividing by k, we find

$$\alpha^{SP}(\mu) = \frac{s - \ell}{k} - \frac{\mu\,\ell}{\lambda \cdot (1 - \mu) \cdot k} \tag{17}$$

This optimum will be valid as long as $0 \leq \alpha^{SP} \leq 1$. To find the bounds, define $\bar{\mu}$ to be the value where $\alpha^{SP}(\bar{\mu}) = 0$. Then, solving for $\bar{\mu}$, we have

$$s - \ell = \frac{\bar{\mu}\ell}{\lambda \cdot (1 - \bar{\mu})}$$

$$\Downarrow$$

$$\bar{\mu} = \frac{\lambda(s - \ell)}{\ell + \lambda(s - \ell)}$$

From eq. (17), we have that α^{SP} is decreasing in μ, so if $\mu > \bar{\mu}$, then eq. (17) will be negative. Similarly, setting $\underline{\mu}$ to be the value where $\alpha^{SP}(\underline{\mu}) = 1$, we have

$$s - \ell - k = \frac{\underline{\mu}\ell}{\lambda \cdot (1 - \underline{\mu})}$$

$$\Downarrow$$

$$\underline{\mu} = \frac{\lambda(s - \ell - k)}{\ell + \lambda(s - \ell - k)}$$

Then for $\mu \leq \underline{\mu}$, we have eq. (17) is greater than 1.

(ii) Recall

$$\begin{aligned}
S &= \mu \cdot \mathcal{W} + (1 - \mu) \cdot \mathcal{K} \qquad \text{(eq. (8))} \\
&= \mu \cdot wLN\left[1 - (1 - \eta)\bar{\alpha}\right] + (1 - \mu) \cdot N\left[\Pi_0 + L\left((s - \ell)\bar{\alpha} - \tfrac{k}{2}\bar{\alpha}^2\right)\right] \qquad \text{(eqs. (8) and (14))} \\
&= N\left[\mu\, wL\left(1 - (1 - \eta)\bar{\alpha}\right) + (1 - \mu)\left(\Pi_0 + L\left((s - \ell)\bar{\alpha} - \tfrac{k}{2}\bar{\alpha}^2\right)\right)\right] \\
&= \underbrace{N\left[\mu\, wL + (1 - \mu)\Pi_0\right]}_{a} + \underbrace{NL\left(-\mu\frac{\ell}{\lambda} + (1 - \mu)(s - \ell)\right)}_{b} \bar{\alpha} + \underbrace{\left(-\frac{(1 - \mu)NLk}{2}\right)}_{\frac{1}{2}\gamma} \bar{\alpha}^2
\end{aligned}$$

where in the linear term we used $w(1 - \eta) = \ell/\lambda$ from (3). This is a quadratic $S = a + b\,\bar{\alpha} + \frac{1}{2}\gamma\,\bar{\alpha}^2$, with $\gamma = -(1 - \mu)\,NLk < 0$ strictly negative. Completing the square around the maximum $x^* = -b/\gamma$:

$$f(x) = a + bx + \tfrac{1}{2}\gamma\,x^2 = f(x^*) + \tfrac{1}{2}\gamma\,(x - x^*)^2,$$

and therefore

$$f(x^*) - f(x) = -\tfrac{1}{2}\gamma\,(x - x^*)^2 = \tfrac{1}{2}|\gamma|\,(x - x^*)^2,$$

where the last equality uses $\gamma \leq 0$. Applying this with $x^* = \alpha^{SP}(\mu)$, $x = \alpha^{NE}$, and $|\gamma| =$

$(1-\mu)\,NLk$:

$$S(\alpha^{SP}) - S(\alpha^{NE}) = \frac{(1-\mu)\,NLk}{2}\,[\alpha^{NE} - \alpha^{SP}(\mu)]^2.$$

(iii) Recall from (9), the wedge decomposes as:

$$\alpha^{NE} - \alpha^{SP}(\mu) = \frac{\ell(1-1/N)}{k} + \frac{\mu\,\ell}{\lambda(1-\mu)\,k}.$$

The first term, $\ell(1-1/N)/k$, is strictly positive because $\ell > 0$ and $N \geq 2$. The second term, $\mu\,\ell/[\lambda(1-\mu)\,k]$, is strictly positive for any $\mu \in (0,1)$ and equals zero at $\mu = 0$. Hence $\alpha^{NE} > \alpha^{SP}(\mu) \geq \alpha^{CO}$ for all $\mu \in [0,1)$.

For worker income: from eq. (15), $\mathcal{W} = wLN[1-(1-\eta)\bar{\alpha}]$ is affine and strictly decreasing in $\bar{\alpha}$ (since $(1-\eta) > 0$ under the maintained assumption $\eta < 1$). Therefore $\alpha^{NE} > \alpha^{CO}$ implies $\mathcal{W}(\alpha^{NE}) < \mathcal{W}(\alpha^{CO})$.

For owner surplus: from eq. (16), $d\mathcal{K}/d\bar{\alpha} = NL[(s-\ell) - k\bar{\alpha}]$, which equals zero at $\bar{\alpha} = (s-\ell)/k = \alpha^{CO}$ and is strictly negative for $\bar{\alpha} > \alpha^{CO}$ (since $k > 0$). Hence $\mathcal{K}$ is strictly decreasing on $(\alpha^{CO}, \alpha^{NE}]$, so $\mathcal{K}(\alpha^{NE}) < \mathcal{K}(\alpha^{CO})$.

Since both $\mathcal{W}$ and $\mathcal{K}$ are strictly lower at the Nash equilibrium than at the cooperative optimum, neither class can be made better off by moving from α^{CO} to α^{NE}: the Nash equilibrium is Pareto dominated.

□

Proof of Corollary 2 *(Sign of the externality).* The wedge formula $\alpha^{NE} - \alpha^{CO} = \ell\,(1-1/N)/k$ is established in Proposition 1(iii) for $\ell = \lambda(1-\eta)w$ and $s > \ell$. The three sign cases follow from the sign of ℓ: $\eta < 1$ gives $\ell > 0$ (positive wedge), $\eta = 1$ gives $\ell = 0$ (zero wedge, and both rates reduce to s/k), and $\eta > 1$ gives $\ell < 0$ (negative wedge). At $\eta = 0$, $\ell = \lambda w$ is maximal, so the wedge $\lambda w(1-1/N)/k$ is maximized. Note that $\eta > 1$ implies $\ell < 0 < s$, so the interior condition $s > \ell$ holds automatically. □

Proof of Proposition 3 *(Worker equity).* With profit-sharing ϵ, aggregate demand satisfies

$$D = A + \lambda\big[\underbrace{wLN(1-(1-\eta)\bar{\alpha})}_{\text{wage income}} + \epsilon\,\Pi\big],$$

where total profit is $\Pi = D - \sum_i C_i = D - NL(w - s\bar{\alpha} + \frac{k}{2}\bar{\alpha}^2)$. Substituting the expression for Π

into the demand equation and collecting D terms on the left-hand side:

$$D(1-\lambda\epsilon) = A + \lambda\, wLN(1-(1-\eta)\bar{\alpha}) - \lambda\epsilon\, NL\big(w - s\bar{\alpha} + \tfrac{k}{2}\bar{\alpha}^2\big)\,.$$

Expanding the right-hand side and grouping by powers of $\bar{\alpha}$:

$$D = \frac{A + \lambda(1-\epsilon)wLN - \ell_\epsilon LN\bar{\alpha} - \lambda\epsilon\frac{k}{2}NL\bar{\alpha}^2}{1-\lambda\epsilon},$$

where $\ell_\epsilon \coloneqq \ell - \lambda\epsilon s = \lambda(1-\eta)w - \lambda\epsilon(w-c)$.

(i) Cooperative optimum. The planner maximizes total profit $\Pi = D - \sum_i C_i$ by choosing the common automation rate α at a symmetric profile. Differentiating D (from the expression above) with respect to α:

$$\frac{\partial D}{\partial \alpha} = \frac{-\ell_\epsilon\, LN - \lambda\epsilon k NL\alpha}{1-\lambda\epsilon}.$$

Differentiating total cost $\sum_i C_i = NL(w - s\alpha + \frac{k}{2}\alpha^2)$:

$$\frac{\partial(\sum_i C_i)}{\partial\alpha} = NL(-s + k\alpha).$$

Setting $\partial\Pi/\partial\alpha = \partial D/\partial\alpha - \partial(\sum_i C_i)/\partial\alpha = 0$ and rearranging:

$$NL(s-k\alpha) = \frac{\ell_\epsilon NL + \lambda\epsilon k NL\alpha}{1-\lambda\epsilon}.$$

Multiplying both sides by $(1-\lambda\epsilon)$:

$$(s-k\alpha)(1-\lambda\epsilon) = \ell_\epsilon + \lambda\epsilon k\alpha.$$

Expanding the left-hand side and collecting α terms:

$$k\alpha\underbrace{\big[(1-\lambda\epsilon)+\lambda\epsilon\big]}_{=1} = s(1-\lambda\epsilon) - \ell_\epsilon.$$

Substituting $\ell_\epsilon = \ell - \lambda\epsilon s$, the right-hand side becomes $s - s\lambda\epsilon - \ell + \lambda\epsilon s = s - \ell$. Hence $k\alpha = s - \ell$, giving $\alpha^{CO}(\epsilon) = (s-\ell)/k$, independent of ϵ.

(ii) Nash equilibrium. Competitive pricing gives $\text{Rev}_i = D/N$. The owner maximizes $(1-\epsilon)\pi_i$; since $(1-\epsilon) > 0$, the FOC reduces to $\partial\pi_i/\partial\alpha_i = 0$. Since D depends on α_i only through $\bar{\alpha}$ (and the $\bar{\alpha}^2$ term), and $\partial\bar{\alpha}/\partial\alpha_i = 1/N$, differentiating $\text{Rev}_i = D/N$ with respect to α_i and evaluating at

a symmetric profile yields

$$\frac{\partial \operatorname{Rev}_i}{\partial \alpha_i} = \frac{1}{N} \cdot \frac{\partial D}{\partial \alpha_i} = \frac{-\ell_\epsilon L - \lambda\epsilon k L \alpha}{N(1-\lambda\epsilon)}.$$

Combining with the cost derivative $\partial C_i/\partial\alpha_i = L(-s+k\alpha)$, the symmetric FOC $\partial \operatorname{Rev}_i/\partial\alpha_i - \partial C_i/\partial\alpha_i = 0$ becomes:

$$s - k\alpha - \frac{\ell_\epsilon + \lambda\epsilon k\alpha}{N(1-\lambda\epsilon)} = 0.$$

Define $N_\epsilon \coloneqq N(1-\lambda\epsilon)+\lambda\epsilon = N - \lambda\epsilon(N-1)$. Collecting the α terms on the left-hand side:

$$k\alpha \underbrace{\left[\frac{N(1-\lambda\epsilon)+\lambda\epsilon}{N(1-\lambda\epsilon)}\right]}_{=N_\epsilon/[N(1-\lambda\epsilon)]} = s - \frac{\ell_\epsilon}{N(1-\lambda\epsilon)}.$$

Multiplying both sides by $N(1-\lambda\epsilon)/N_\epsilon$:

$$k\alpha = \frac{s\,N(1-\lambda\epsilon)-\ell_\epsilon}{N_\epsilon}.$$

Substituting $\ell_\epsilon = \ell - \lambda\epsilon s$ into the numerator:

$$s\,N(1-\lambda\epsilon) - (\ell - \lambda\epsilon s) = s\big[N(1-\lambda\epsilon)+\lambda\epsilon\big] - \ell = s\,N_\epsilon - \ell.$$

Hence $k\alpha = (sN_\epsilon - \ell)/N_\epsilon = s - \ell/N_\epsilon$, giving $\alpha^{NE}(\epsilon) = (s-\ell/N_\epsilon)/k$.

(iii) Wedge. Subtracting the cooperative optimum from the Nash rate:

$$\alpha^{NE}(\epsilon) - \alpha^{CO} = \frac{s-\ell/N_\epsilon}{k} - \frac{s-\ell}{k} = \frac{\ell - \ell/N_\epsilon}{k} = \frac{\ell(N_\epsilon - 1)}{k\,N_\epsilon}.$$

Since $N_\epsilon = N - \lambda\epsilon(N-1)$, we have $N_\epsilon - 1 = (N-1) - \lambda\epsilon(N-1) = (N-1)(1-\lambda\epsilon)$, so the wedge becomes

$$W(\epsilon) = \frac{\ell(N-1)(1-\lambda\epsilon)}{k\,N_\epsilon}.$$

To show the wedge is strictly decreasing in ϵ, write $W = \ell(N-1)(1-\lambda\epsilon)/[k(N-\lambda\epsilon(N-1))]$ and apply the quotient rule. The numerator is $f(\epsilon) = 1-\lambda\epsilon$ with $f' = -\lambda$, and the denominator is $\tilde{g}(\epsilon) = N - \lambda\epsilon(N-1)$ with $\tilde{g}' = -\lambda(N-1)$. By the quotient rule:

$$\frac{d}{d\epsilon}\left(\frac{f}{\tilde{g}}\right) = \frac{f'\tilde{g} - f\tilde{g}'}{\tilde{g}^2} = \frac{-\lambda[N-\lambda\epsilon(N-1)] + \lambda(N-1)(1-\lambda\epsilon)}{N_\epsilon^2} = \frac{-\lambda}{N_\epsilon^2} < 0.$$

Hence $\partial W/\partial\epsilon = -\lambda\ell(N-1)/(kN_\epsilon^2) < 0$: the wedge is strictly decreasing. Setting $W=0$ requires $1-\lambda\epsilon = 0$, i.e., $\epsilon = 1/\lambda$, which satisfies $\epsilon \geq 1$ whenever $\lambda \leq 1$. □

Proof of Corollary 3 *(Voluntary profit-sharing).* Firm i chooses $\epsilon_i \in [0,1]$ to maximize retained profit $(1-\epsilon_i)\pi_i$. Differentiating with respect to ϵ_i:

$$\frac{d}{d\epsilon_i}\big[(1-\epsilon_i)\pi_i\big] = -\pi_i + (1-\epsilon_i)\frac{\partial \pi_i}{\partial \epsilon_i}.$$

When firm i shares ϵ_i of its profit, its workers receive $\epsilon_i\pi_i$ and spend $\lambda\epsilon_i\pi_i$ in the sector. Firm i captures $1/N$ of the resulting demand increase, so $\partial\pi_i/\partial\epsilon_i = \lambda\pi_i/N$. Evaluating at $\epsilon_i = 0$:

$$-\pi_i + \frac{\lambda\pi_i}{N} = \pi_i\left(\frac{\lambda}{N} - 1\right) < 0,$$

since $\lambda \le 1$ and $N \ge 2$. The derivative is negative at the left endpoint, so $\epsilon_i = 0$ is optimal regardless of the other firms' choices. ☐

Proof of Proposition 4. Partition the N firms into a coalition $\mathcal{M}$ of size M and a fringe of $N-M$ non-members. Let

$$\alpha^F := \frac{1}{N-M} \cdot \sum_{i \notin \mathcal{M}} \alpha_i \tag{18}$$

be the average automation rate of fringe firms. At a symmetric profile where all coalition members choose α^M, the average automation rate is

$$\bar{\alpha} = \frac{M\alpha^M + (N-M)\alpha^F}{N}.$$

From (6), a coalition member i's profit is

$$\pi_i = \Pi_0 + L\left[\alpha_i\left(s - \frac{\ell}{N}\right) - \frac{k}{2}\alpha_i^2 - \frac{\ell}{N}\sum_{j\neq i}\alpha_j\right].$$

Then

$$\begin{aligned}
\pi_i &= \Pi_0 + L\left[\alpha_i\left(s - \frac{\ell}{N}\right) - \frac{k}{2}\alpha_i^2 - \frac{\ell}{N}\left(\sum_{j\in\mathcal{M}\setminus\{i\}}\alpha_j + \sum_{j\notin\mathcal{M}}\alpha_j\right)\right] \\
&= \Pi_0 + L\left[\alpha_i\left(s - \frac{\ell}{N}\right) - \frac{k}{2}\alpha_i^2 - \frac{\ell}{N}\left(\sum_{j\in\mathcal{M}\setminus\{i\}}\alpha_j + (N-M)\alpha^F\right)\right].
\end{aligned}$$

Under a symmetric strategy, where all firms in the coalition follow the same strategy $\alpha_i = \alpha^M$, we have

$$\begin{aligned}\pi_i &= \Pi_0 + L\left[\alpha^M\left(s - \frac{\ell}{N}\right) - \frac{k}{2}\left(\alpha^M\right)^2 - \frac{\ell}{N}\left((M-1)\alpha^M + (N-M)\alpha^F\right)\right]\\ &= \Pi_0 + L\left[\alpha^M\left(s - \frac{\ell M}{N}\right) - \frac{k}{2}\left(\alpha^M\right)^2 + \frac{\ell}{N}(N-M)\alpha^F\right]\\ &= \Pi_0 + L\frac{\ell(N-M)\alpha^F}{N} + L\left[\alpha^M\left(s - \frac{\ell M}{N}\right) - \frac{k}{2}\left(\alpha^M\right)^2\right]\end{aligned}$$

The coalition chooses α^M to maximize $\sum_{i\in\mathcal{M}}\pi_i$. At the symmetric equilibrium, $\sum_{i\in\mathcal{M}}\pi_i = M\pi_i$, so maximizing the individual firm's profit in the coalition also maximizes the coalition profits. Taking the derivative, we find

$$\frac{\partial\pi_i}{\partial\alpha^M} = L\left(s - \frac{M\ell}{N} - k\alpha^M\right) = L\left(s - \frac{\ell}{N} - k\alpha^M - \frac{(M-1)\ell}{N}\right)$$

The additional term $-(M-1)\ell/N$ (relative to eq. (13)) arises because each coalition member internalizes the demand loss its automation imposes on the other $M-1$ members, each of whom loses $\ell L/N$ in revenue. Setting the first-order condition to zero:

$$\alpha^M = \frac{s - M\ell/N}{k},$$

clamped to $[0,1]$. At $M=1$, this reduces to $\alpha^{NE} = (s-\ell/N)/k$. At $M=N$, it reduces to $\alpha^{CO} = (s-\ell)/k$.

When both α^M and α^{CO} are interior, the residual wedge is

$$\alpha^M - \alpha^{CO} = \frac{s - M\ell/N}{k} - \frac{s-\ell}{k} = \frac{\ell(1-M/N)}{k},$$

which is strictly positive for $M<N$ and zero only at $M=N$. □

Proof of Proposition 5 *(Pigouvian automation tax).*

(i) Under the tax, firm i's profit from (6) becomes

$$\pi_i = \Pi_0 + L\left[\alpha_i\left(s - \tau - \frac{\ell}{N}\right) - \frac{k}{2}\alpha_i^2 - \frac{\ell}{N}\sum_{j\neq i}\alpha_j\right].$$

The first-order condition is $s - \tau - \ell/N - k\alpha_i = 0$, giving $\alpha^{NE}(\tau) = (s-\tau-\ell/N)/k$. Setting $\alpha^{NE}(\tau) = \alpha^{CO} = (s-\ell)/k$ yields $\tau^* = \ell - \ell/N = \ell(1-1/N)$.

(ii) At $\tau = \tau^*$, all firms choose α^{CO}. Per-firm profit is $\pi^{\text{tax}} = \pi(\alpha^{CO}) - \tau^* L\alpha^{CO} = \pi^{CO} - \tau^* L\alpha^{CO}$. Total tax revenue is $\tau^* LN\alpha^{CO}$; rebated equally, each firm receives $\tau^* L\alpha^{CO}$, restoring profit to π^{CO}.

□

Proof of Proposition 6 *(AI productivity widens the over-automation wedge).*

(i) At a symmetric profile $\alpha_i = \alpha$ with $\phi > 1$, firm i's revenue is $\text{Rev}_i = DY_i/(N\bar{Y})$. The firm's first-order condition equates the marginal benefit of automation to its marginal cost: $\partial \text{Rev}_i / \partial\alpha_i + sL - kL\alpha_i = 0$, where sL is the per-unit cost saving from replacing a worker with AI and $kL\alpha_i$ is the marginal integration friction. Substituting the marginal revenue from (10) and rearranging, the symmetric first-order condition is

$$k\alpha = s - \frac{\ell}{N} + \frac{D(\alpha)\,(\phi-1)(N-1)}{N^2[1+(\phi-1)\alpha]\,L}.$$

Define $\text{LHS}(\alpha) = k\alpha$ and let $\text{RHS}(\alpha)$ denote the right-hand side. The left-hand side is strictly increasing (slope k). The right-hand side is strictly decreasing: the market-share term has numerator proportional to $D(\alpha) = A + \lambda wLN - \ell LN\alpha$ (decreasing in α) and denominator factor $1 + (\phi-1)\alpha$ (increasing in α). Hence LHS = RHS has a unique solution.

To show $\alpha^{NE}(\phi) > \alpha^{NE}(1)$, evaluate both sides at the baseline equilibrium $\alpha = \alpha^{NE}(1) = (s - \ell/N)/k$: LHS $= s - \ell/N$, while RHS $= s - \ell/N +$ (positive market-share term) $>$ LHS. Since LHS is increasing and RHS is decreasing, the unique crossing must occur at $\alpha^{NE}(\phi) > \alpha^{NE}(1)$.

(ii) The cooperative planner maximizes $\sum_i \pi_i = D - \sum_i C_i$. Total revenue equals aggregate demand D regardless of how output is allocated across firms: expenditure D is pinned down by (2), so reallocating output shares does not change total revenue. The planner's first-order condition therefore depends only on costs:

$$\frac{\partial(\sum_i \pi_i)}{\partial\alpha_i} = -\ell L + sL - kL\alpha_i,$$

which is independent of ϕ. Setting to zero gives $\alpha^{CO}(\phi) = (s-\ell)/k = \alpha^{CO}(1)$.

For the generalized planner: $S(\mu) = \mu\,\mathcal{W} + (1-\mu)\,\mathcal{K}$. Worker income $\mathcal{W} = wLN[1-(1-\eta)\bar{\alpha}]$ does not depend on ϕ. Owner surplus $\mathcal{K} = D - \sum_i C_i$ at symmetric profiles, and neither D (2) nor C_i (1) depends on ϕ. Hence $S(\mu)$ is ϕ-invariant at every symmetric $\bar{\alpha}$, and $\alpha^{SP}(\mu;\phi) =$

$\alpha^{SP}(\mu; 1)$ for all μ.

(iii) Combining (i) and (ii): $\alpha^{NE}(\phi) > \alpha^{NE}(1)$ while $\alpha^{SP}(\mu; \phi) = \alpha^{SP}(\mu; 1)$ for all μ, so the wedge $\alpha^{NE}(\phi) - \alpha^{SP}(\mu; \phi)$ is strictly larger than $\alpha^{NE}(1) - \alpha^{SP}(\mu; 1)$ for every μ. Since $\alpha^{NE}(\phi)$ is increasing in ϕ (by the same LHS/RHS argument with a larger market-share term), the wedge is strictly increasing in ϕ.

□

Proof of Proposition 7 *(Endogenous entry, frictionless benchmark).* We assume $k = 0$ (frictionless case), $\lambda = 1$ (full recycling), $0 < \kappa < A$ (entry is costly but the market is viable for at least one firm), and $\ell > s$ (so that $N^* > 1$ by Corollary 1).

Step 1: the profit schedule is strictly decreasing on $\mathbb{N}$. Let $m = \lfloor N^* \rfloor$, so $m \le N^* < m + 1$. For $N \le N^*$, corollary 1 gives $\alpha = 0$, so by (5) with $\lambda = 1$: $\Pi^*(N) = A/N$, which is strictly decreasing. For $N > N^*$, full automation is dominant (corollary 1), and per-firm profit drops by $\Delta = L(\ell - s) > 0$, giving $\Pi^*(N) = A/N - \Delta$, also strictly decreasing. At the crossing: $\Pi^*(m) = A/m > A/(m + 1) > A/(m + 1) - \Delta = \Pi^*(m + 1)$. Hence Π^* is strictly decreasing on $\mathbb{N}$.

Step 2: existence and uniqueness of N^{FE}. Since $\kappa < A$, $\Pi^*(1) = A > \kappa$. Since $\Delta > 0$, $\Pi^*(N) \to -\Delta < 0$ as $N \to \infty$. The set $\mathcal{S} = \{N \in \mathbb{N} : \Pi^*(N) \ge \kappa\}$ is therefore nonempty and finite. Let $N^{FE} = \max \mathcal{S}$. By strict monotonicity, $\Pi^*(N^{FE}) \ge \kappa$ and $\Pi^*(N^{FE} + 1) < \kappa$, so N^{FE} satisfies (11) and is unique.

Step 3: characterization by regime. We determine N^{FE} by checking which integers are in $\mathcal{S}$.

On the no-automation branch ($N \le m$): $\Pi^*(N) = A/N \ge \kappa$ iff $N \le A/\kappa$. So the viable integers on this branch are $\{1, \ldots, \min(\lfloor A/\kappa \rfloor, m)\}$.

On the full-automation branch ($N \ge m + 1$): $\Pi^*(N) = A/N - \Delta \ge \kappa$ iff $N \le A/(\kappa + \Delta)$. The smallest integer on this branch is $m + 1$, so viable integers exist iff $m + 1 \le A/(\kappa + \Delta)$, i.e. $\kappa + \Delta \le A/(m + 1)$. When nonempty, the viable set is $\{m + 1, \ldots, \lfloor A/(\kappa + \Delta) \rfloor\}$.

Since N^{FE} is the largest viable integer overall:

Case (i): (low entry cost) $\kappa + \Delta \le A/(m + 1)$. The full-automation branch contains viable integers up to $\lfloor A/(\kappa + \Delta) \rfloor \ge m + 1$. Since $\kappa + \Delta > \kappa$, we have $A/(\kappa + \Delta) < A/\kappa$, so the full-automation branch's maximum does not exceed the no-automation branch's maximum in absolute terms, but it exceeds m (the cap on the no-automation branch). Hence $N^{FE} = \lfloor A/(\kappa + \Delta) \rfloor \ge m + 1 > N^*$, and every firm fully automates. See fig. 4(a).

Case (ii): (intermediate entry cost) $\kappa + \Delta > A/(m + 1)$ *and* $\kappa < A/m$. The first condition means $A/(m + 1) - \Delta < \kappa$, so $\Pi^*(m + 1) = A/(m + 1) - \Delta < \kappa$. Since Π^* is decreasing, no integer above N^* is viable. The second condition gives $A/\kappa > m$, so $\lfloor A/\kappa \rfloor \geq m$. Hence the largest viable integer on the no-automation branch is m, and $N^{FE} = m$. Profit is $\Pi^*(m) = A/m > \kappa$ (strict, since $A/\kappa > m$ implies $A/m > \kappa$). In this case since $N^{FE} = m < N^*$, no firm automates. See fig. 4(b).

Case (iii): (high entry cost) $\kappa \geq A/m$. Then $A/\kappa \leq m$, so $\lfloor A/\kappa \rfloor \leq m$. Since $\kappa + \Delta > \kappa \geq A/m > A/(m+1)$, no integer on the full-automation branch is viable (same argument as case (ii)). Hence $N^{FE} = \lfloor A/\kappa \rfloor$. Since $N^{FE} = \lfloor A/\kappa \rfloor \leq m \leq N^*$, no firm automates. See fig. 4(c).

These three cases exhaust all $\kappa \in (0, A)$. When $\kappa > A$, then $N^{FE} = 0$. ☐

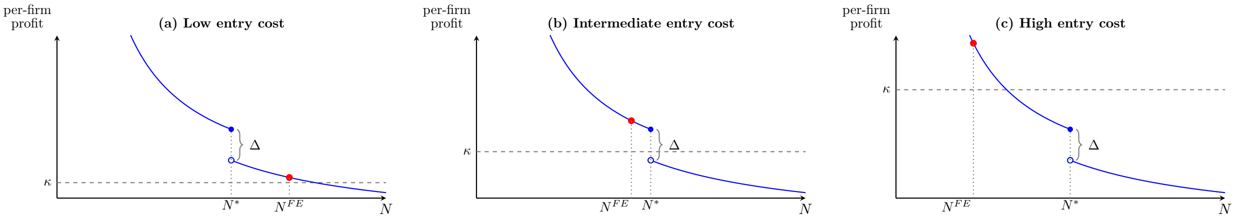

Figure 4: Per-firm profit as a function of the number of firms in the frictionless benchmark ($k = 0$, $\lambda = 1$). In each panel, profit follows A/N for $N \leq N^*$ (no automation) and drops discretely by $\Delta = L(\ell - s)$ at N^* when full automation becomes dominant. The red dot marks the free-entry equilibrium N^{FE}. (a) Low entry cost: N^{FE} lies above N^* and all firms automate. (b) Intermediate entry cost: $N^{FE} = \lfloor N^* \rfloor$; the threat of automation deters the marginal entrant, sustaining positive profits without any automation occurring. (c) High entry cost: N^{FE} falls well below N^*; entry costs alone limit competition and the automation threshold is never approached.

***Proof of Proposition 8** (Endogenous entry with convex costs).* First, we show that the per-firm profit is decreasing in N. From eq. (5), we can write

$$\pi^{NE}(N) = A/N + C + g(N) \tag{19}$$

where

$$C := (\lambda - 1)wL \tag{20}$$

$$g(N) := L\,\alpha^{NE}(N)\left[(s - \ell) - \tfrac{k}{2}\alpha^{NE}(N)\right] \tag{21}$$

Since A/N is strictly decreasing on $\mathbb{N}$, it suffices to show that g is weakly decreasing for $N \in \mathbb{N}$. First, note that $g(\cdot)$ is decreasing in α^{NE}

$$\frac{dg}{d\alpha^{NE}} = (s - \ell) - k\,\alpha^{NE} \tag{22}$$

By assumption $\ell > s$, so the derivative is always negative. Then, by Proposition 1, we have

$$\alpha^{NE} = \begin{cases} 0 & \text{if } N < \frac{\ell}{s} \\ \frac{s-\ell/N}{k} & \text{if } \frac{\ell}{s} \leq N \leq \frac{\ell}{s-k} \\ 1 & \text{if } \frac{\ell}{s-k} < N \end{cases} \tag{23}$$

So α^{NE} is non-decreasing in N. Thus $g(\alpha^{NE}(N))$ is non-increasing in N.

Since A/N is decreasing in N, and $g\left(\alpha^{NE}(N)\right)$ is non-increasing in N, so $\pi^{NE}(N) = A/N + C + g(N)$ is strictly decreasing in N, i.e., $\pi^{NE}(N) > \pi^{NE}(N+1)$.

Now, let

$$S := \{N \in \mathbb{N} : \pi^{NE}(N) \geq \kappa\} \tag{24}$$

Since $\pi^{NE}(1) = \Pi_0(1) > \kappa$ by assumption, the set S is non-empty. To see that S is finite, we need to show that $\lim_{N\to\infty} \pi^{NE} < \kappa$. First,

$$\alpha_\infty := \lim_{N\to\infty} \alpha^{NE}(N) = \min\left(1, \frac{s}{k}\right) \tag{25}$$

So

$$\lim_{N\to\infty} g(N) = L\,\alpha_\infty\left[(s-\ell) - \tfrac{k}{2}\alpha_\infty\right] < 0 \tag{26}$$

because $s < \ell$ (by assumption). Thus

$$\lim_{N\to\infty} \pi^{NE}(N) = \lim_{N\to\infty} [A/N + C + g(N)] = (\lambda - 1)wL + \lim_{N\to\infty} g(N) < 0 \tag{27}$$

Since $C = (\lambda-1)wL < 0$. Since $\lim_{N\to\infty} \pi^{NE} < 0 < \kappa$, the set S is finite. Let

$$N^{FE} := \max\left(N \in S\right) \tag{28}$$

Then $\pi^{NE}(N^{FE}) \geq \kappa$ and $\pi^{NE}(N^{FE}+1) < \kappa$, so N^{FE} satisfies (11). If $N^{FE} > N^*$, then $\alpha^{NE}(N^{FE}) = \min\left((s - \ell/N^{FE})/k, 1\right) > 0$; since $\ell > s$ implies $\alpha^{CO} = \max\{0, (s-\ell)/k\} = 0$, we have $\alpha^{NE}(N^{FE}) > \alpha^{CO}$, so over-automation persists. □

***Proof of Proposition* 9** *(Endogenous wages).* The symmetric equilibrium is a fixed point: $\bar{\alpha}$ such that $\bar{\alpha} = \alpha^{NE}(w(\bar{\alpha}))$.

(i) At a symmetric rate $\alpha_i = \alpha$ for all i, we have $\bar{\alpha} = \alpha$ and, from (5),

$$\pi_i \;=\; \Pi_0 + L\big(s\,\alpha - \ell\,\alpha - \tfrac{k}{2}\alpha^2\big)\,.$$

Aggregate profit is $\sum_i \pi_i = N\,\pi_i$. Differentiating with respect to α and dividing by NL yields the planner's per-firm marginal benefit of automation:

$$g(\alpha) \coloneqq s(\alpha) - \ell(\alpha) - k\alpha = w(\alpha) - c - \lambda(1-\eta)w(\alpha) - k\alpha = w(\alpha)\big[1 - \lambda(1-\eta)\big] - c - k\alpha,$$

where we used $s = w - c$ and $\ell = \lambda(1-\eta)w$. By contrast, each firm's profit from (6) depends on $\bar{\alpha}$ only through its own revenue share ℓ/N. Differentiating with respect to α_i and dividing by L gives the private marginal benefit:

$$h(\alpha) \coloneqq s(\alpha) - \frac{\ell(\alpha)}{N} - k\alpha = w(\alpha)\left[1 - \frac{\lambda(1-\eta)}{N}\right] - c - k\alpha.$$

Differentiating, $g'(\alpha) = w'(\alpha)[1 - \lambda(1-\eta)] - k$; the first term is weakly negative and $k > 0$, so $g' < 0$. The same argument applies to h (whose coefficient $1 - \lambda(1-\eta)/N > 0$ for $N \geq 2$), so both are strictly decreasing. Note that $g(\alpha) = h(\alpha) - \ell(\alpha)(1 - 1/N)$ for all α. At the NE fixed point, $h(\alpha^{NE}) = 0$ by definition, so

$$g(\alpha^{NE}) = 0 - \ell(\alpha^{NE})(1 - 1/N) = -\ell(\alpha^{NE})(1 - 1/N) < 0.$$

Since g is strictly decreasing and $g(\alpha^{CO}) = 0$ (the planner's optimality condition), the inequality $g(\alpha^{NE}) < 0 = g(\alpha^{CO})$ implies $\alpha^{CO} < \alpha^{NE}$.

(ii) From (7), the threshold is $N^*(w) = \lambda(1-\eta)w/(w-c)$. Applying the quotient rule:

$$\frac{\partial N^*}{\partial w} = \frac{\lambda(1-\eta)(w-c) - \lambda(1-\eta)w}{(w-c)^2} = \frac{-\lambda(1-\eta)c}{(w-c)^2} < 0,$$

so N^* is strictly decreasing in w. Since $w'(\bar{\alpha}) \leq 0$ by assumption, $w(\bar{\alpha}) \leq w(0)$ for all $\bar{\alpha} \geq 0$. Combining: $N^*(w(\bar{\alpha})) \geq N^*(w(0))$, with strict inequality whenever $w(\bar{\alpha}) < w(0)$. □

Corollary 4 (Generalized planner under wage adjustment)**.** *Under the conditions of Proposition 9, let* $\mu \in [0,1)$ *and define* $g_\mu(\alpha) \coloneqq s(w(\alpha)) - \ell(w(\alpha))\big[1 + \mu/(\lambda(1-\mu))\big] - k\alpha$. *If* g_μ *is strictly decreasing, then* $\alpha^{NE} > \alpha^{SP}(\mu)$.

Proof sketch. The argument mirrors part (i) of Proposition 9, replacing the cooperative planner's marginal benefit g with the μ-planner's. The μ-planner maximizes $S(\mu) = \mu\,\mathcal{W} + (1-\mu)\,\mathcal{K}$ over a common rate α, taking wages as given. From the proof of Proposition 2, the first-order condition is

$$(1-\mu)\big[s(\alpha) - \ell(\alpha) - k\alpha\big] = \frac{\mu\,\ell(\alpha)}{\lambda},$$

which can be rewritten as $g_\mu(\alpha) = 0$. This function relates to the private marginal benefit $h(\alpha) = s(\alpha) - \ell(\alpha)/N - k\alpha$ from the proof of Proposition 9 by

$$g_\mu(\alpha) = h(\alpha) - \ell(\alpha)\left(1 - \frac{1}{N} + \frac{\mu}{\lambda(1-\mu)}\right).$$

At the Nash equilibrium, $h(\alpha^{NE}) = 0$, so

$$g_\mu(\alpha^{NE}) = -\ell(\alpha^{NE})\left(1 - \frac{1}{N} + \frac{\mu}{\lambda(1-\mu)}\right) < 0,$$

since $\ell > 0$, $N \geq 2$, and $\mu > 0$. At the planner's fixed point, $g_\mu(\alpha^{SP}(\mu)) = 0$ by definition. If g_μ is strictly decreasing, the inequality $g_\mu(\alpha^{NE}) < 0 = g_\mu(\alpha^{SP})$ implies $\alpha^{SP}(\mu) < \alpha^{NE}$, exactly as in the $\mu = 0$ case.

It remains to verify monotonicity. Differentiating:

$$g'_\mu(\alpha) = w'(\alpha)\left[1 - (1-\eta)\left(\lambda + \frac{\mu}{1-\mu}\right)\right] - k.$$

Define the bracket as $C_\mu := 1 - (1-\eta)[\lambda + \mu/(1-\mu)]$. At $\mu = 0$, $C_0 = 1 - \lambda(1-\eta) \geq 0$, which is the coefficient used in the proof of Proposition 9; since $w' \leq 0$, the product $w'C_0 \leq 0$ and $g'_0 < 0$ follows immediately. For $\mu > 0$, C_μ decreases. As long as $C_\mu \geq 0$, the same argument applies; this holds for all $\mu \leq \bar{\mu} := [1 - (1-\eta)\lambda]/[2 - \eta - (1-\eta)\lambda]$ (approximately 0.48 at $\lambda = 0.5$, $\eta = 0.30$). When $C_\mu < 0$, the product $w'C_\mu \geq 0$, so $g'_\mu < 0$ requires $k > |w'(\alpha)| \cdot |C_\mu|$: integration frictions must dominate wage sensitivity.

As a numerical illustration, consider $w(\bar{\alpha}) = 1 - 0.5\,\bar{\alpha}$, $c = 0.30$, $\lambda = 0.5$, $\eta = 0.30$, $k = 1$, $N = 7$, and $\mu = 0.3$. The equilibrium rates are $\alpha^{NE} \approx 0.44$, $\alpha^{CO} \approx 0.26$, and $\alpha^{SP}(0.3) \approx 0.04$, confirming $\alpha^{SP} < \alpha^{CO} < \alpha^{NE}$. The distributional premium is substantial: the μ-planner would reduce automation to near zero, well below the cooperative optimum that already lies far below the Nash rate. □

Proof of Proposition 10 *(Capital income recycling).* We derive the result for general $k \geq 0$; the proposition's two parts follow from the $k = 0$ specialization.

Aggregate demand is $D = A + \lambda wLN - \ell LN\bar{\alpha} + \hat{\eta}\Pi$, where total profit is $\Pi = D - NL(w - s\bar{\alpha}) - \frac{k}{2}L\sum_j \alpha_j^2$. Substituting and solving for D:

$$D(1-\hat{\eta}) = A + (\lambda - \hat{\eta})wLN - \ell_{\hat{\eta}}\, LN\bar{\alpha} - \tfrac{\hat{\eta}k}{2}L\textstyle\sum_j \alpha_j^2,$$

with $\ell_{\hat{\eta}} = \ell - \hat{\eta}s$. When $k = 0$ this gives (12). Since $N\bar{\alpha} = \alpha_i + \sum_{j \neq i} \alpha_j$, differentiating:

$$\frac{\partial D}{\partial \alpha_i} = -\frac{\ell_{\hat{\eta}}\, L}{1 - \hat{\eta}} - \frac{\hat{\eta} k L}{1 - \hat{\eta}}\, \alpha_i.$$

Revenue $\mathrm{Rev}_i = D/N$ gives $\partial \mathrm{Rev}_i / \partial \alpha_i = -\ell_{\hat{\eta}} L/[N(1-\hat{\eta})] - \hat{\eta} k L \alpha_i/[N(1-\hat{\eta})]$. Firm i's marginal profit, including the direct cost saving sL and marginal friction $kL\alpha_i$, is

$$\frac{\partial \pi_i}{\partial \alpha_i} = L\left(s - \frac{\ell_{\hat{\eta}}}{N(1-\hat{\eta})} - k\alpha_i \cdot \frac{\hat{N}}{N(1-\hat{\eta})}\right),$$

where $\hat{N} \coloneqq N(1-\hat{\eta}) + \hat{\eta}$. This depends only on α_i, so the equilibrium is in strictly dominant strategies.

Part (i): $k = 0$. The marginal profit reduces to $L(s - \ell_{\hat{\eta}}/[N(1-\hat{\eta})])$, a constant independent of α_i. This is positive if and only if $sN(1-\hat{\eta}) > \ell_{\hat{\eta}}$, i.e. $N > \frac{\ell_{\hat{\eta}}}{s(1-\hat{\eta})} = N_{\hat{\eta}}$. Thus full automation is strictly dominant when $N > N_{\hat{\eta}}$ and no automation is strictly dominant when $N < N_{\hat{\eta}}$, reproducing the structure of Corollary 1.

Part (ii). $\ell_{\hat{\eta}} = \ell - \hat{\eta}s \leq 0$ if and only if $\hat{\eta} \geq \ell/s$.

Extension to $k > 0$. Setting the first-order condition to zero:

$$\alpha^{NE} = \frac{N(1-\hat{\eta})}{k\hat{N}}\left(s - \frac{\ell_{\hat{\eta}}}{N(1-\hat{\eta})}\right) = \frac{s\hat{N} - \ell}{k\hat{N}} = \frac{s - \ell/\hat{N}}{k}.$$

This is positive if and only if $\hat{N} > N^*$, reproducing Proposition 1 with N replaced by $\hat{N}$. At a symmetric profile ($\alpha_j = \alpha$ for all j), total profit is $\Pi = \frac{N}{1-\hat{\eta}}[\Pi_0 + L(s-\ell)\alpha - \frac{k}{2}L\alpha^2]$. The $1/(1-\hat{\eta})$ multiplier scales the objective without changing the optimizer, so $\alpha^{CO} = \max\{0,\ (s-\ell)/k\}$, the same as in Proposition 1. $\square$