The Price of Intelligence: How Should Socially-minded Firms Price and Deploy AI?*

Nils H. Lehr
International Monetary Fund

Pascual Restrepo
Yale University

September 3, 2025

Abstract

Leading AI firms claim to prioritize social welfare. *How should firms with a social mandate price and deploy AI?* We derive pricing formulas that depart from profit maximization by incorporating incentives to improve welfare and reduce labor disruptions. Using US data, we evaluate several scenarios. A *welfarist firm* that values both profit and welfare should price closer to marginal cost, as efficiency gains outweigh distributional concerns. A *conservative firm* focused on labor-market stability should price above the profit-maximizing level in the short run, especially when its AI may displace low-income workers. Overall, socially minded firms face a trade-off between expanding access to AI and the resulting loss in profits and labor market risks.

*The views expressed in this paper are our own and do not necessarily reflect those of the IMF, its Executive Board, or its Management. Restrepo is part of the Anthropic Economic Advisory Council.

Artificial Intelligence (AI) promises to transform the economy, raising new questions about how firms should price, deploy, and manage this technology. Leading AI firms present themselves as socially responsible entities. They claim a dual mandate: to generate profits for shareholders while enhancing social welfare and mitigating risks. *OpenAI* adopted a capped-profit model. Investors earn returns up to a fixed multiple, after which the organization prioritizes its mission to "benefit all of humanity" by "building safe and beneficial AGI and helping create broadly distributed benefits."[1] *Anthropic* declares to "make decisions that maximize positive outcomes for humanity in the long run".[2] Both companies claim to have been conservative in deploying more advanced models and capabilities, aiming to manage societal risks—such as economic displacement—while giving the labor market time to adjust.

How should firms with a social mandate price and deploy AI? Is a commitment to maximizing shareholder returns the best way to promote welfare? Should they expand access by pricing below profit-maximizing levels? Or should they deploy AI slowly to mitigate labor market risks?

This paper addresses these questions by providing optimal-price formulas for socially minded AI firms. The formulas extend *Lerner's Rule*, which says that profit-maximizing firms should set

$$\frac{P - MC}{P} = \frac{1}{\varepsilon},$$

with ε the demand elasticity. Optimal pricing is given by a *Modified Lerner Rule*

$$\frac{P - MC}{P} = \frac{\mathcal{M}}{\varepsilon},$$

where $\mathcal{M}$ summarizes the motives of a socially minded firm. We derive the formulas in a general equilibrium environment where a tech firm has a monopoly over an AI capable of replicating human skills. The deployment of this AI reduces production costs but disrupts labor markets for workers with skills that are substitutable. The AI firm prioritizes profits, broader social welfare, and minimizing labor market disruptions.

The optimal deployment strategy balances four distinct considerations:

- Profit motives push towards $\mathcal{M} = 1$, as in the traditional Lerner Rule.

[1] See https://openai.com/index/openai-elon-musk/.
[2] https://www.anthropic.com/company

- Aggregate efficiency considerations push towards $\mathcal{M} = 0$, or marginal-cost pricing. This achieves the level of AI production and access that maximizes the size of the pie.
- These aggregate benefits are weighed against distributional considerations, which capture who benefits the most from AI. These can be positive or negative, depending on whether AI substitutes for high- or low-income workers.
- Finally, the incentive to minimize labor market disruptions pushes for higher values of $\mathcal{M}$ that can exceed one in the short run but not in the long run. This motive calls for a gradual deployment path, with the firm acting conservatively. This is because the cost of disrupting the labor market is higher in the short run, while workers adjust.

The formulas highlight a tension between expanding access to AI (to maximize aggregate efficiency) and the resulting short-term loss in profits and labor market risks.

We then present an exploration of the formulas, using US data. We compute the optimal deployment path and prices of an AI capable of replacing human labor in each of 525 detailed jobs. For each job, we imagine our tech firm develops an AI capable of replacing labor at 50% of the cost and ask how a socially minded firm should price and deploy such AI.

We report optimal plans for firms that value welfare and minimizing disruptions to varying degrees. A *welfarist firm* that values profits and welfare should price closer to marginal cost. This is because for all jobs considered, efficiency gains outweigh distributional concerns by a wide margin, since losses do not concentrate among low-income workers. On the other hand, a *conservative firm* focused solely on balancing profits with labor-market stability should price above the profit-maximizing level in the short run. A firm that values welfare and stability equally should price close to the profit-maximizing level in the short run and closer to marginal cost in the long run.

We conclude that the most pro-social course of action for AI firms with considerable market power is to refrain from exploiting it. A socially minded AI firm should price closer to marginal cost in an effort to broaden access, with the only possible exception being the very short run, when stability concerns are most significant. This conclusion contradicts recommendations to tax AI and automation technologies to mitigate their adverse effects on the labor market. What these recommendations miss is that AI firms can have considerable market power. If we worry that an AI can have sizable impacts on prices and wages and is controlled by a small number of firms, we must accept the possibility that these firms wield considerable market power and would limit output to bid up prices. This exercise of market power already protects workers from the substituting effects of AI at the expense of consumers. Further increasing the price of AI through taxes or self-regulation would have an adverse first-order impact on consumers, whose access to AI is already limited, with only modest protective benefits for workers.

We conclude the paper with extensions that explore the robustness of this conclusion. For example, we demonstrate that the incentive for socially minded AI firms to price closer to marginal cost becomes stronger when a progressive tax system is in place, providing some redistribution and insurance for workers, or when its AI does not substitute for workers but instead creates value by introducing new products. Conversely, we demonstrate that distributional and stability concerns become more significant when there is increased competition among AI suppliers.

Literature This paper contributes to the long-standing debate on the social responsibilities of firms. Following Friedman (1970), the traditional view holds that a firm's sole obligation is to maximize shareholder value. Leading AI companies explicitly reject this view by adopting mission statements that emphasize societal welfare, long-term human outcomes, and labor stability. This paper explores how such objectives should alter their pricing strategies.

This paper also contributes to the growing literature on optimal policy responses to AI and automation. A strand examines the optimal taxation of automation technologies, motivated either by distributional concerns (Guerreiro, Rebelo and Teles, 2021; Donald, 2022; Thuemmel, 2023; Costinot and Werning, 2022; Lehr and Restrepo, 2024; Bond and Kremens, 2025) or efficiency considerations (Acemoglu, Manera and Restrepo, 2020; Beraja and Zorzi, 2022). Our work relates to this literature in that socially responsible AI firms partially internalize distributional concerns by curbing the scale of AI deployment—much like how a tax on automation can reduce its use and mitigate inequality. However, a key distinction is that, in the models studied in the literature, it is always optimal to tax technologies that worsen inequality (assuming the set of fiscal tools is limited). This result relies on the assumption of an efficient baseline economy, where the cost of a small tax is second-order, while the distributional gains are first-order. In contrast, our setting begins with an inefficient allocation due to market power, resulting in insufficient AI production. In this context, expanding the use of AI yields first-order efficiency gains, which must be balanced against concerns regarding distributional and labor market stability.

A third related literature study the optimal deployment of AI accounting for existential risks (Jones, 2024, 2025) and social risks that can be learned over time or via testing (Acemoglu and Lensman, 2024; Guerreiro, Rebelo and Teles, 2023). Our work abstracts from these risks and focuses exclusively on the question of how firms should deploy well-aligned or narrow AIs that carry no existential risks.

Finally, our paper contributes to a growing empirical literature exploring how AI could disrupt labor markets by measuring the capabilities of AI (Webb, 2020; Brynjolfsson, Mitchell and Rock, 2018; Felten, Raj and Seamans, 2021, 2023; Eloundou et al., 2023; Handa et al., 2025) and studying the deployment of AI and Large Language Models in specific contexts (Peng et al., 2023; Brynjolfsson, Li and Raymond, 2023; Noy and Zhang, 2023). These papers show that AI can substitute for human labor in various domains at a fraction of the cost and with minimal input from expert human workers. We use some of the estimates from these papers in our numerical exploration.

1 Model of labor-replacing AI

This section outlines a general model of how AI affects wages, prices, and households' welfare. We focus on AI technology capable of replicating human skills or inputs in some areas of the economy. Examples include the use of AI systems to automate tasks such as radiology, copywriting, journalism, customer service, and driving. These are all domains where AI systems can be trained to replicate human input. We also assume that the technology is sufficiently advanced to operate autonomously and without requiring input from workers. In our discussion section, we extend our theory to account for the possibility that AI is used for novel applications beyond replicating human input.

1.1 The Economy

The economy flows in continuous time t. There is a discrete set of commodities $j \in \mathcal{J}$ and skills or labor inputs $s \in \mathcal{S}$. Commodity $j = 0$ serves as the numeraire.

The economy is populated by a mass ϵ of financiers and a mass 1 of regular households (identified with the superscript h). Financiers own firms and no labor endowments. They consume the numeraire good and make consumption and saving decisions to maximize

$$u \equiv \int_0^\infty e^{-\rho t} \, c_{0t} \, dt \;\; \text{st:} \;\; \dot{a}_t = r_t a_t + \pi_t - c_{0t}.$$

Regular household h is endowed with a vector of skills or labor inputs $n_t^h = (n_{st}^h)_{s\in\mathcal{S}}$ that can change over time. They consume commodity bundles $c_t^h = (c_{jt}^h)_{j\in\mathcal{J}}$ and maximize

$$u^h \equiv \int_0^\infty e^{-\rho t} \, u(c_t^h) \, dt \;\; \text{st:} \;\; \dot{a}_t^h = r_t a_t^h + w_t \cdot n_t^h - p_t \cdot c_t^h \;\; \text{and} \;\; a_t^h \in \mathcal{R}.$$

Here $p_t = (p_{jt})_{j\in\mathcal{J}}$ is the price of commodities at time t (with $p_{0t} = 1$) and $w_t = (w_{st})_{s\in\mathcal{G}}$ are wages, with household wages given by $w_t^h = w_t \cdot n_t^h$. The term $a_t^h \in \mathcal{R}$ captures potential constraints, assumed independent of prices.

AI can replicate labor input in a subset $\mathcal{A}$ of $\mathcal{S}$. The quantity of s input is

$$\ell_{st} = \begin{cases} \displaystyle\int_h n_{st}^h \, dh \; + \; q_{st} \;\; \text{for} \;\; s \in \mathcal{A} \\[2ex] \displaystyle\int_h n_{st}^h \, dh \;\; \text{otherwise.} \end{cases}$$

Here, q_{st} represents units of AI-generated output, assumed to be indistinguishable from that of workers.

To produce AI-generated output, the AI firm uses $1/\psi_{st}$ units of computing resources, where ψ_{st} denotes the efficiency of algorithms reproducing input s. Computing resources, denoted as x_t, are produced at a one-to-one rate from the numeraire commodity. Feasibility requires

$$\sum_{s\in\mathcal{A}} \psi_{st} \, q_{st} \;\leq\; x_t,$$

so that the consumption of computational resources by AI does not exceed supply.

Commodities y are produced using labor (or AI) ℓ. Plans $y = (y_{jt})_{j\in\mathcal{J},t}$ and $\ell = (\ell_{st})_{s\in\mathcal{S},t}$ with

$$F(y, \ell) \leq 0$$

can be produced. F has constant returns to scale and is operated competitively. Feasibility requires

$$c^{\omega} + \int_h c_{0t}^h \, dh + x_t \leq y_{0t} \text{ and}$$

$$\int_h c_{jt}^h \, dh \leq y_{jt} \text{ otherwise}$$

so that consumption of commodities does not exceed production.

Equilibrium: we are interested in an equilibrium where the AI company sets a feasible choice of q_{st} and x_t anticipating the effects of its actions on prices, profits, and the economy.

Given the choices of q_{st} and x_t, the equilibrium is defined in a standard way. It is given by a set of prices $\{r_t, p_t, w_t\}$, consumption plans $\{c_t^h, c_{0t}\}$, asset positions $\{a_t^h, a_t\}$, and production plans y, ℓ such that consumers maximize utility subject to their flow-budget constraint and asset restrictions, competitive firms maximize profits from operating F taking prices as given, commodity markets clear, and the asset market clears. Equilibrium profits for the AI-producing firm at time t are

$$\pi_t = \sum_{s \in \mathcal{A}} (w_{st} - \psi_{st}) \, q_{st}$$

To derive our formulas, we do not need to solve for the equilibrium explicitly. It suffices that financiers set the interest rate $r_t = ho$ and determine the discount factor used by firms.

The objective of socially-oriented AI firms: The AI firm operates under three objectives: profit maximization, social welfare, and minimizing labor market disruptions. Its objective function is

$$V = \text{PDV } \pi_t \; + \int_h \mu^h \, u^h \, dh \; + \; \lambda \int_{h: w_t^h < \bar{w}^h} \text{PDV } \frac{w_t^h}{\bar{w}^h} \, dh.$$

The first term captures profit maximization motives.

The second term captures welfare considerations in a reduced-form way. Here μ^h is the value the firm attaches to increasing the income of household h. The μ^h's differ across households, reflecting distributional considerations. As in standard welfare functions, the firm attaches greater weight to poor households than richer ones. Investor welfare is already accounted for in profits, so we do not include it again to avoid double-counting.

The third term captures the objective of minimizing labor-market disruptions created by AI, with a weight of λ. The AI firm penalizes labor-market losses incurred by exposed households, computed as the percent decline in labor income of household h relative to its initial status quo of $\bar{w}^h$. These penalties represent various considerations. Firms may adopt the principle that reducing people's wages below their status quo level is undesirable, either because people are particularly averse to wage losses or because the firm adopts a *conservative* stance when judging its labor-market impact that regards these deviations as unfair (as in Corden, 1974). Penalties may also capture strategic considerations, with the firm minimizing disruptions to reduce discontent. In our formulation, the firm penalizes all wage losses, without accounting for indirect benefits via reduced product prices. AI firms may attach greater weight to wage losses because people are more sensitive or responsive to their labor-market outcomes, either because these are more salient (benefits from reduced product prices are "out of sight; out of mind") or because they derive status from their high wages.[3] To summarize, AI firms want to avoid major shifts in the way labor markets operate, with the status and wages of different jobs falling in ways that may be perceived as unfair or arbitrary by workers.

To simplify the exposition, we derive formulas assuming a quasi-linear aggregator of the form

$$u(c_t^h) = c_{0t}^h + \sum_j u_j(c_{jt}^h).$$

We also assume the equilibrium is such that all households h consume $c_{0t}^h > 0$ at all times.

To understand firm incentives, consider how a deviation in plans $\{\delta q_{st}\}$ affects its objective:

$$\begin{aligned}\delta V = \int_0^t e^{-\rho t} \Bigg\{ &(1-\mu) \sum_{s\in\mathcal{A}} \left(q_{st}\ \delta w_{st} + \left(w_{st} - \psi_{st}\right) \delta q_{st}\right) \\ &+ \mu \sum_{s\in\mathcal{A}} \left(w_{st} - \psi_{st}\right) \delta q_{st} + \mu \int_h g^h \sum_s n_{st}^h\ \delta w_{st}\ dh + \lambda \int_{h: w_t^h < \bar{w}^h} \frac{1}{\bar{w}^h} \sum_s n_{st}^h\ \delta w_{st}\ dh \Bigg\}\ dt. \end{aligned} \tag{1}$$

Here, $\mu = \int_h \mu^h\ dh$ is the average welfare weight across households and $g^h = \mu^h/\mu - 1$ are the normalized weights. By construction, $\int_h g^h\ dh = 0$ and the sign of g^h represents distributional motives.

[3] In our formulation, the AI firm penalizes a reduction in wages in *percent terms*, so that a reduction in wages of $10,000 receives a higher penalty if experienced by low-income workers.

The first term in the right of (1) represents *profit motives*. We assume $1 > \mu$ so that the firm has an incentive to maximize profits.

The second term represents *efficiency motives*. Because the firm cares about welfare, it has an incentive to produce efficiently, increasing quantities until prices equal marginal cost, $P = MC$.

The third term represents *distributional motives*. These call for restricting the quantity of AI produced if it competes against poor households. This motive is weighed against efficiency considerations.

The last term represents *conservative motives*. These receive a weight λ and capture the value of minimizing the labor-market disruptions generated by AI. These are different from standard distributional motives in that the AI firm is concerned about disrupting the labor market of both rich and poor households, all of whom experience some wage pressure due to the deployment of AI. In writing this, we assumed all households are exposed to AI, in the sense that $n^h_{st} > 0$ for at least some $s \in \mathcal{A}$.

The firm optimally balances these motives to ensure $\delta V = 0$. This implies:

Proposition 1. *In interior equilibria of the quasi-linear case, the socially minded firm produces q_{st} until*

$$\mathcal{L}_{st} = \sum_{s'} \left((1-\mu) \frac{q_{s't}\ w_{s't}}{q_{st}\ w_{st}} + \mu \int_h g^h \frac{n^h_{s't}\ w_{s't}}{q_{st}\ w_{st}}\, dh + \lambda \int_{h: w^h_t < \bar{w}^h} \frac{1}{\bar{w}^h} \frac{n^h_{s't}\ w_{s't}}{q_{st}\ w_{st}}\, dh \right) \frac{1}{\varepsilon_{s'st}} \tag{2}$$

where $\varepsilon_{ss't} \equiv -\frac{\partial \ln q_{st}}{\partial ln w_{s't}}$ is the cross demand elasticity between s and s' and $q_{s't} = 0$ for $s' \notin \mathcal{A}$.

Proof. In an interior equilibrium where $q_{st} > 0$, any deviation in q_{st} must yield $\delta V = 0$. Setting $\delta V = 0$ in (1) and rearranging yields (2). □

To develop intuition, assume the cross-demand elasticity is 0 for $s \neq s'$. The own demand elasticity (ε_{st} for $s = s'$) is strictly positive and always remains in the formula. Let's also consider first a profit-maximizing firm, by setting $\mu = \lambda = 0$. The formula says that a profit-maximizing AI firm should restrict quantities until its Lerner index $\mathcal{L} \equiv (P - MC)/P$ satisfies Lerner's Rule

$$\mathcal{L}_{st} = \frac{1}{\varepsilon_{st}}, \tag{3}$$

where $\varepsilon_{st} \geq 0$ is the (negative) of the demand elasticity of s at time t.

For a socially minded firm, optimal pricing satisfies a Modified Lerner Rule

$$\mathcal{L}_{st} = \frac{\mathcal{M}}{\varepsilon_{st}},$$

where

$$\mathcal{M} \equiv 1 - \mu \ + \ \mu \int_h g^h \, \frac{n_{st}^h}{q_{st}} \, dh \ + \ \lambda \int_{h:w_t^h < \bar{w}^h} \frac{1}{\bar{w}^h} \, \frac{n_{st}^h}{q_{st}} \, dh.$$

The adjustment term differs from 1 and summarizes the different firm considerations:

- The "1" is the usual profit maximization term.
- The "$-\mu$" pushes towards lower markups and higher quantities. This term reflects the firm's desire to increase access to AI, thereby raising aggregate efficiency at the expense of investors.
- The term "$\mu \int_h g^h \frac{n_{st}^h}{q_{st}} dh$" has ambiguous sign. It is positive when AI competes more intensely against poor households. In this case, AI deepens existing inequalities, causing a socially minded firm to restrict its use by charging higher prices. The term can be negative if AI competes more intensely against rich households. In this case, the use of AI reduces underlying inequalities, causing socially minded firms to lower prices and increase quantities.
- The term "$\lambda \int_{h:w_t^h < \bar{w}^h} \frac{1}{\bar{w}^h} \frac{n_{st}^h}{q_{st}} dh$" is always positive and reduces quantities of AI produced. This captures the AI firm's incentive to minimize labor market disruptions. This incentive to curb the use of AI is stronger when it competes against poor segments of the labor market, since a reduction in wages of a given amount is more costly in proportional terms for low-wage households.

The formula serves to illustrate several scenarios. For a *utilitarian* AI firm that cares about profits and aggregate efficiency but has no distributional or conservative inclinations ($\mu > 0, g^h = 0, \lambda = 0$), optimal prices satisfy

$$\mathcal{L}_{st} = (1 - \mu) \, \frac{1}{\varepsilon_{st}}.$$

These prices are below the profit-maximizing level and closer to marginal-cost pricing. For a *welfarist* AI firm that cares about profits, welfare, and distributional issues, but has no distributional or conservative inclinations ($\mu > 0, g^h \neq 0, \lambda = 0$), optimal prices satisfy

$$\mathcal{L}_{st} = \left(1 - \mu \ + \ \mu \int_h g^h \frac{n^h_{st}}{q_{st}} \, dh\right) \frac{1}{\varepsilon_{st}}.$$

For a *conservative* AI firm that cares about minimizing labor market disruptions but not about welfare per se ($\mu = 0, \lambda > 0$), optimal prices satisfy

$$\mathcal{L}_{st} = \left(1 \ + \ \lambda \int_{h: w^h_t < \bar{w}^h} \frac{1}{\bar{w}^h} \frac{n^h_{st}}{q_{st}} \, dh \right) \frac{1}{\varepsilon_{st}},$$

which exceed the profit-maximizing level.

In the general case with cross effects ($\varepsilon_{ss't} \neq 0$ for $s \neq s'$), the formula accounts for equilibrium price effects on all workers and revenue from other AI products. For example, when $\mu = \lambda = 0$, we recover the standard multi-product Lerner formula, which takes into account how increasing the quantity supplied of one good affects demand for other AI products sold by the firm.

1.2 A Tractable Example of Equilibrium with Socially-Minded Firms

In general, the equilibrium of the model is given by (i) a choice of quantities and prices by the AI firm that satisfy the Modified Lerner's rule, (ii) a vector of commodity prices, and (iii) production and consumption plans that maximize households' utility and firms' profits (for firms producing commodities y). The characterization of the equilibrium is generally complicated, as the residual demand for AI depends on how skills are combined into goods, the demand for these goods by households, and the supply of skills.

In this sub-section, we characterize the full equilibrium of the model in an example economy with the following features:

(a) Each commodity is produced linearly using a commodity-specific skill, with the skill associated with the numeraire commodity not in $\mathcal{A}$.

(b) The utility function is $u_s(c_s) = \gamma_s^{1/\sigma_s} \, \frac{c_s^{1-1/\sigma_s}}{1-1/\sigma_s}$, with $\sigma_s > 1$, so that the demand for each commodity has a constant elasticity σ_s.

(c) Households reallocate labor away from disrupted skills at a rate $\alpha > 0$. This implies

$$n_{st}^h = \bar{n}_s^h \, e^{-\alpha\, t} \;\text{ and } n_{st} = \bar{n}_s \, e^{-\alpha\, t} \;\text{ for }\; s \in \mathcal{A}.$$

Here, $\{\bar{n}_s^h\}$ and $\bar{n}_s$ denote pre-AI quantities of labor input in skill s.

(d) AI is productive enough to justify deployment and ensure an interior equilibrium. This implies

$$1 - \mu \int_h g^h \, \frac{\bar{n}_s^h}{\bar{n}_s} \, dh \;-\; \lambda \int_h \frac{1}{\bar{w}^h} \, \frac{\bar{n}_s^h}{\bar{n}_s} \, dh \;>\; \frac{\psi_{st}}{\bar{w}_s},$$

where ψ_{st} is the marginal cost of the AI firm and $\bar{w}_s = \gamma_s \, \bar{n}_s^{-1/\sigma_s}$ the pre-AI price of skill s.

In this economy, the quantity and price of AI for each skill are determined independently.

Proposition 2. *In an economy where (a)–(d) hold, equilibrium prices and quantities of skills in $\mathcal{A}$, are uniquely determined by two equations. The supply curve, obtained by rearranging (2):*

$$1 - \frac{\psi_{st}}{w_{st}} = \left(1 \;-\; \mu \;+\; \mu \int_h g^h \, \frac{\bar{n}_s^h}{q_{st}} \, e^{-\alpha\, t} \, dh \;+\; \lambda \int_h \frac{1}{\bar{w}^h} \, \frac{\bar{n}_s^h}{q_{st}} \, e^{-\alpha\, t} \, dh \right) \frac{q_{st}}{q_{st} + \bar{n}_s \, e^{-\alpha\, t}} \, \frac{1}{\sigma_s} \tag{4}$$

and the demand curve, obtained from consumer demand:

$$w_{st} = \gamma_s \, (q_{st} + \bar{n}_s \, e^{-\alpha\, t})^{-1/\sigma_s} \tag{5}$$

Proof. Equation (4) follows from the formula in Proposition 1, using the fact that in this economy, the elasticity of demand for AI (accounting for worker production) exceeds σ_s and is given by

$$\varepsilon_{st} \;=\; \frac{q_{st} + \bar{n}_s e^{-\alpha t}}{q_{st}} \, \sigma_s.$$

The demand curve in (5) is derived by equating the marginal rate of substitution for commodity s (relative to the numeraire) to its price w_s.

Note that in this economy, there are no complementarities across jobs. As a result, $w_t^h < \bar{w}^h$ for all households with $n_{st}^h > 0$ for some $s \in \mathcal{A}$. This is why the cost of disruption sums over all h. □

The proposition provides formulas for computing the full deployment path of an AI that substitutes for skill s. The supply and demand curve pin down quantities, prices, and markups charged in equilibrium by socially minded AI firms. Figure 1 depicts the supply and demand curves, assuming the distributional motive is positive. Condition (d) ensures the curves intersect at a unique $q_{st} > 0$.

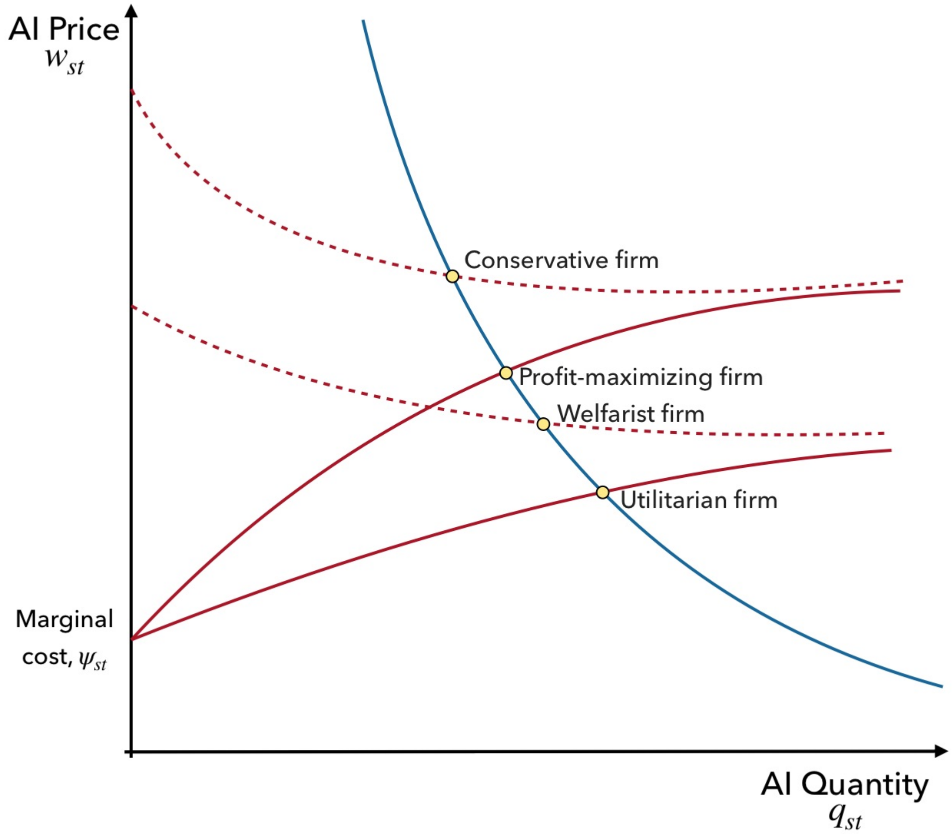

Figure 1: Equilibrium Supply and Demand for AI

Notes: The figure shows the demand and supply curves for an AI that substitutes for skill s. The supply curve and equilibrium points are shown for a profit-maximizing firm, a utilitarian firm, a welfarist firm (assuming the distributional motive is positive), and a conservative firm.

The supply curve for a profit-maximizing firm is upward sloping: as the quantity of AI produced increases, the residual demand curve becomes more inelastic, leading to higher markups. The utilitarian firm supply curve is shifted to the right, reflecting incentives to charge lower markups to increase access and aggregate efficiency. The supply curves of welfarist and conservative firms are shifted upward, reflecting incentives to restrict quantities and mitigate the harmful distributional or labor-market impacts of AI.

The figure also shows that the distributional motives of a welfare-maximizing firm or the stability motives of a conservative firm vanish as quantities increase. This is why the supply curve of a welfarist firm converges to the utilitarian one and the supply curve of a conservative firm converges to the profit-maximizing one. This force can be so strong as to render the supply curve of these firms downward sloping—a distinct possibility shown in the Figure. From equation (4), this is the case if

$$\mu \int_h g^h \frac{\bar{n}_s^h}{\bar{n}_s}\, dh \;+\; \lambda \int_h \frac{1}{\bar{w}^h} \frac{\bar{n}_s^h}{\bar{n}_s}\, dh > 1 - \mu,$$

so that distributional and labor-market stability concerns are dominant.

To understand why distributional and labor-market stability concerns vanish, return to equation (1), describing the effects of changes in quantities produced on the objective of the AI firm. The firm balances three objectives: profits, aggregate efficiency, and distributional and stability considerations. The equation indicates that profit and efficiency motives are directly proportional to the quantity of AI used. Increasing the quantity of AI by 1% leads to a larger profit and efficiency increase when the AI is widely used. However, distributional and stability concerns do not scale with quantities. Increasing quantities produced by 1% reduces wages of exposed groups by at most $(1/\sigma_s) \times 1\%$—an effect that remains bounded as the use of AI deepens. For this reason, socially minded firms prioritize efficiency and profit motives as the use of AI becomes widespread.[4]

The formulas in the proposition also highlight two new economic mechanisms introduced by labor reallocation. First, the formulas show that distributional and labor-market stability motives vanish over time as workers reallocate. This force calls for a gradual and backloaded deployment plan, where AI firms first curb quantities and set higher prices to shield exposed workers from disruptions and give them time to adjust. Over time, firms lower prices and expand quantities, as workers slowly reallocate away from exposed skills or sectors of the labor market.

Second, the reallocation of labor away from exposed skills eases competition, making the residual demand faced by the AI firm more inelastic over time. This allows firms to set higher markups in the long run, leading to a more front-loaded deployment plan.

The net effect of these forces over time on markups and pricing is ambiguous. For a pure profit-maximizing firm, the second effect is the only one present, and we would expect markups to increase over time as the AI firm becomes the sole supplier of skills in $\mathcal{A}$. For a conservative firm, the second effect might dominate, leading to markups that decrease in time.

[4]The same logic is explored in Costinot and Werning (2022). Their paper derives formulas for optimal taxes that balance aggregate efficiency with distributional considerations. As here, the cost of distorting trade or the use of automation technology scales with quantities, which calls for smaller taxes on trade and technology as globalization deepens and the use of automation technology becomes widespread.

2 Scenarios for AI Transitions with Socially-Minded Firms

We now turn to a numerical exploration of our formulas. We focus on the example economy in Proposition 2 and ask the hypothetical question:

> *Imagine a firm develops an AI capable of replicating skill s at a fixed fraction of its current cost. How should socially minded firms deploy and price this technology*?

The formulas in the proposition demonstrate how to calculate the optimal deployment path for any such AI. By focusing on these hypothetical AIs, we avoid the more challenging question of determining which specific skills are most likely to be automated in the near term.

For our application, we focus on a firm that operates in the US economy and map skills to 525 detailed occupations from the 2017–2021 American Community Survey. For each occupation, we compute the optimal deployment plan of an AI capable of replacing labor inputs in said occupation.

For the model parameters, we set a reallocation rate $\alpha = 4\%$ per year, in line with estimates from our previous work (Lehr and Restrepo, 2024). We also set $\sigma_s = 3$, which is a commonly used value for the elasticity of substitution between differentiated goods, as the ones produced by different skills in our model (see, for example, Broda and Weinstein, 2006).[5]

The data inputs needed for our calculations and appearing in the formulas from Proposition 2 are computed as follows:

- We let h denote the set of people at different percentiles of the US income distribution, assumed to have the same relative weight g^h.[6]

- We take $\bar{w}_s$ as the average hourly wage across occupations from 2017–2021 ACS. For each percentile, we then measure $\bar{y}_s^h$ as their income from occupation s and define

$$\bar{n}_s^h = \frac{\bar{y}_s^h}{\bar{w}_s},$$

  as the effective hours worked by households from the hth percentile in occupation s.

[5] A related object is the elasticity of substitution between college and non-college labor, with estimates ranging from 1.4 (as in Katz and Murphy, 1992) to 4 (as in Bils, Kaymak and Wu, 2024). For broad occupations, Burstein, Morales and Vogel (2019) estimate an elasticity of substitution of 2.1. We use a larger value since our occupational groups are finer.

[6] In defining these percentiles, we sort individuals based on household income per person. This is computed as total household income divided by the number of adults plus a half times the number of children. This approach accounts for intra-household income sharing, assigning children a weight of 0.5 times that of an adult.

- We let $\bar{w}_h$ denote the average labor income of people in percentile h.

- We compute total labor input in s as $\bar{n}_s = \sum_h \bar{n}_s^h$ and calibrate γ_s to match $\bar{w}_s = \gamma_s \ \bar{n}_s^{-1/\sigma_s}$.

Finally, we assume $\psi_s = .5 \ \bar{w}_s$, so that AI can replicate human labor at 50% the cost.[7] The rationale for this choice is as follows. In our model, an AI substituting for skill s and sold at a standard markup $\sigma_s \ / \ (\sigma_s - 1)$ above marginal cost raises output per worker from 1 to

$$1 + \frac{q_{st}}{\bar{n}_s} = \left(\frac{\psi_{st}}{\bar{w}_s} \ \frac{\sigma_s}{\sigma_s - 1} \right)^{-\sigma_s} = 2.4.$$

This 2.4-fold increase in output per worker matches the upper end of available empirical estimates. For example, Noy and Zhang (2023) estimate a twofold increase in (quality-adjusted) output per worker in writing tasks and Brynjolfsson, Li and Raymond (2023) estimate a 1.15 increase in customer service.

In the analysis, we contrast the optimal deployment plans of various firms. We consider:

- a pure profit maximizer ($\mu = g_h = \lambda = 0$);

- a utilitarian firm ($\mu > 0$, $g_h = \lambda = 0$);

- a welfarist firm ($\mu > 0, g_h \neq 0, \lambda = 0$);

- a conservative firm ($\mu = g_h = 0, \lambda > 0$);

- a multi-objective firm ($\mu > 0, g_h \neq 0, \lambda > 0$).

In the relevant scenarios, we set $\mu = 0.5$ and use the welfare weights g^h reported in Lockwood and Weinzierl (2016), inferred from the progressivity of the US tax system. This assumes that the welfare weights of the AI firm align with those that the US political system places on households at different percentiles of the income distribution. Our value for μ implies the firm is willing to trade 1 dollar of profit for 2 dollars of value for the economy as a whole. The values for welfare weight g^h are shown in Figure 2. The values imply that the firm is willing to give \$ 1 of profits to increase incomes by \$1.90 at the bottom of the income distribution and \$ 3 at the top.

[7] Variable costs include the computational resources needed to run the AI and effectively replicate human input in skills s, plus any residual costs associated with integration, prompting, or inspection of the AI output. Replicating human input can require multiple calls to these models, explaining why ψ_{st} can vary across jobs. The variable computational and energy costs of using AI are significant and have increased as AI companies train larger and more complex models with higher inference costs.

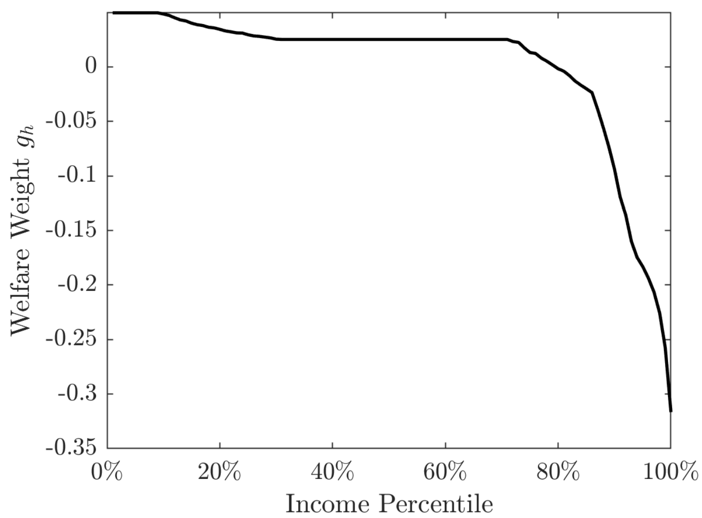

Figure 2: Welfare Weights Across the Income Distribution

Notes: The figure reports welfare weights g_h by income percentile. These are obtained as the welfare weights that rationalize the progressiveness of the US tax system, and are based on Figure 1 in Lockwood and Weinzierl (2016). We directly use the reported weights at specific income percentiles and interpolate between them to span the income distribution.

Finally, in the relevant scenarios, we rescaled λ by the average wage in the economy to ensure that all terms have an equal scale and set $\lambda = 0.5$. This implies that the firm is willing to reduce profits by \$ 1 if it raises wages for the average displaced worker by \$ 2. These scenarios are meant to clarify how AI firms may act if they pursue a broader set of social objectives; of course, we do not know what is in the minds or hearts of their CEOs or how they will weigh different considerations in practice.

2.1 Equilibrium markups and AI deployment plans

Figure 3 reports equilibrium markups $(w_{st} - \psi_{st})/\psi_{st}$ for firms with different objectives at three time horizons. Panel A shows markups on impact ($t = 0$), Panel B for the short run ($t = 5$ years), and Panel C for the long run ($t = 100$ years). The figures sort the 525 detailed occupations by their average base wage $\bar{w}_s$ in the horizontal axis. The movement along the curves shows how markups vary across occupations hypothetically replaced by AI as we move from low-pay to high-pay roles.

As a benchmark, consider a pure profit-maximizing firm, in black. Markups for this firm at $t = 0$ are around 32% and constant, since we assume a common productivity improvement across all skills. As expected from our discussion of 2, markups rise at longer time horizons, reflecting reduced competition from workers as they reallocate to other jobs. In the long run, the AI firm

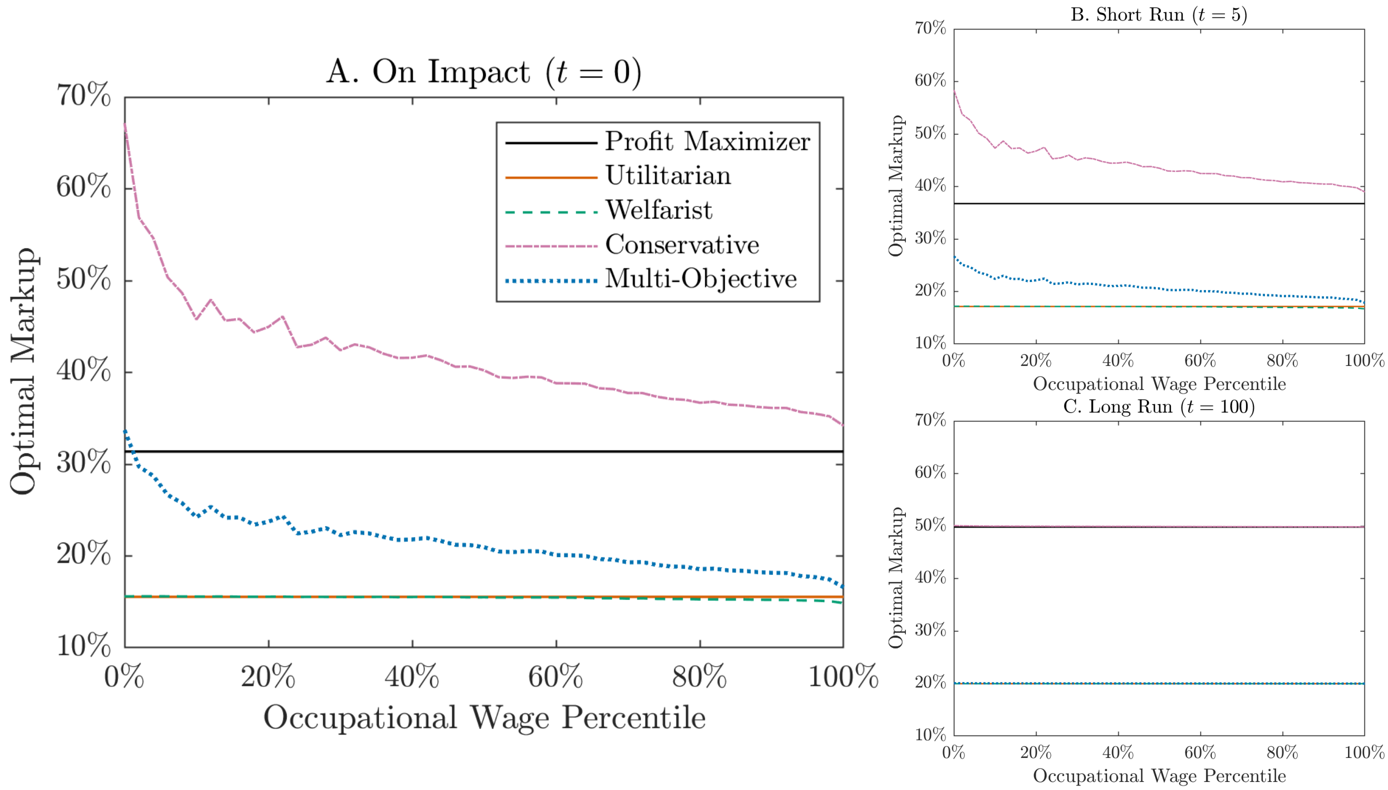

Figure 3: Equilibrium Markups in the Short and Long Run

Notes: Panel A reports equilibrium markups on impact ($t = 0$) for AIs capable of automating different occupations (ranked by wage in the horizontal axis). The curves are smoothed by binning occupations into 50 quantiles and reporting the average within each bin. Each panel shows five curves, one for each type of firm. Panels B and C report the same curves after 5 and 100 years.

becomes the sole supplier of skill s and charges a markup of $\sigma_s/(\sigma_s - 1) = 50\%$ across the board.

The utilitarian firm, in orange, charges lower markups than the profit-maximizing firm, about 15% at $t = 0$ and converging to 20% in the long run. This is because the utilitarian firm has an incentive to lower prices below the profit-maximizing level to expand access and increase aggregate efficiency.

The welfarist firm, in dashed green, prioritizes both aggregate efficiency and distributional concerns. The latter have a tiny impact on equilibrium markups at the bottom. Relative to the utilitarian firm, distributional concerns call for a 0.1 percentage point higher markup at the bottom, thereby redistributing resources towards low-income households in low-paying jobs. Distributional considerations have a modest impact on markups at the top, lowering them (relative to the utilitarian firm) by a full percentage point (from 15% to 14%). These results suggest that distributional considerations play a small role. From the viewpoint of a welfarist firm, the concern of maximizing access dominates and leads to markups that are less than half of what a profit-maximizing firm would charge.

The conservative firm, in solid purple, balances profits against labor-market stability. This firm ends up charging prices above the profit-maximizing level to minimize its labor market impact. This concern is particularly pronounced for AIs that automate low-wage occupations, as these generate more substantial labor-market disruptions. For this reason, equilibrium markups are higher at the bottom, for AIs that automate low-paying occupations. In the long run, stability concerns vanish and the firm stops behaving conservatively to focus entirely on profit maximization.

Finally, the dashed blue line presents markups for a multi-objective firm, which balances profit, efficiency, redistribution, and stability concerns. This firm charges a 33% markup on AIs competing against low-paying workers and a 15% markup on AIs that substitute labor in high-paying occupations. In the long run, distributional and stability considerations fade as workers reallocate, and the AI firm converges to a common 20% markup that balances broader access with its profit motive.

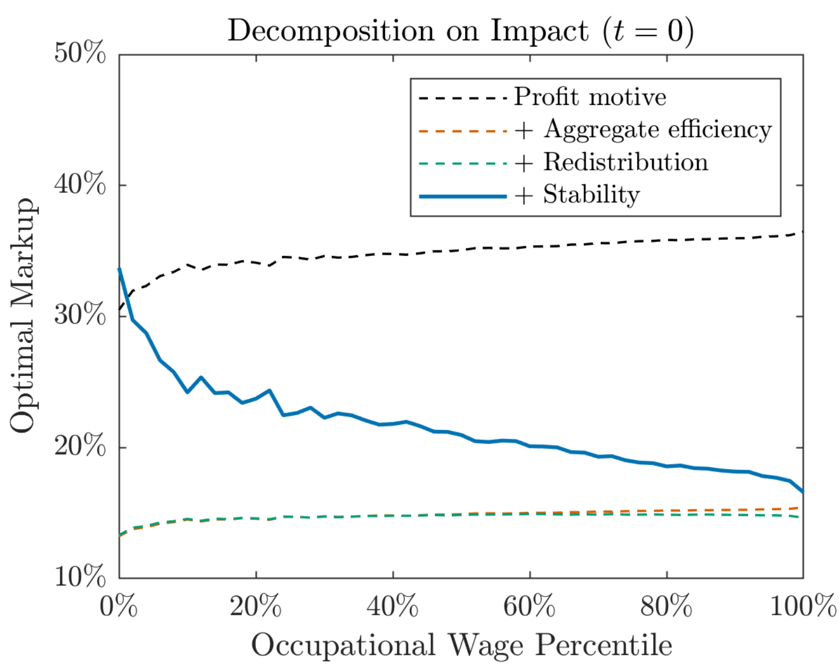

Figure 4: Decomposition of Motives Driving Markups Charged by Multi-Objective Firm

Notes: The figure decomposes equilibrium markups charged by a multi-objective firm on impact ($t = 0$) for AIs capable of automating different occupations (in the horizontal axis). The solid blue line depicts the equilibrium markup. The black dotted line represents the contribution of profit-maximizing motives. The orange dotted line adds the contribution of aggregate efficiency considerations. The green line takes into account distributional considerations. The gap between this and the solid blue reflects labor-market stability considerations. The curves are smoothed by binning occupations into 50 quantiles and reporting the average within each bin. Each panel shows five curves, one for each type of firm. Panels B and C report the same curves after 5 and 100 years.

Figure 4 decomposes the role of each motive for the multi-objective firm. The dashed black line represents the contribution of profit motives, which push for high markups, especially for high-paying jobs facing less competition from workers. The orange line incorporates aggregate efficiency considerations, which call for uniformly lower markups to balance profit against increased access. The green line accounts for distributional considerations, which have no impact at the bottom, and calls for lower markups at the top. Finally, the blue line takes into account wage stability concerns, which call for curbing quantities and raising prices for AIs, especially those that replace low-wage jobs.

Figure 5 complements the results by reporting equilibrium quantities. We plot the increase in quantities relative to their baseline levels before AI implementation. AI-produced quantities range from one to four times the baseline level. The utilitarian firm generates the maximum increase in AI usage, while the conservative and profit-maximizing firms restrict quantities the most.

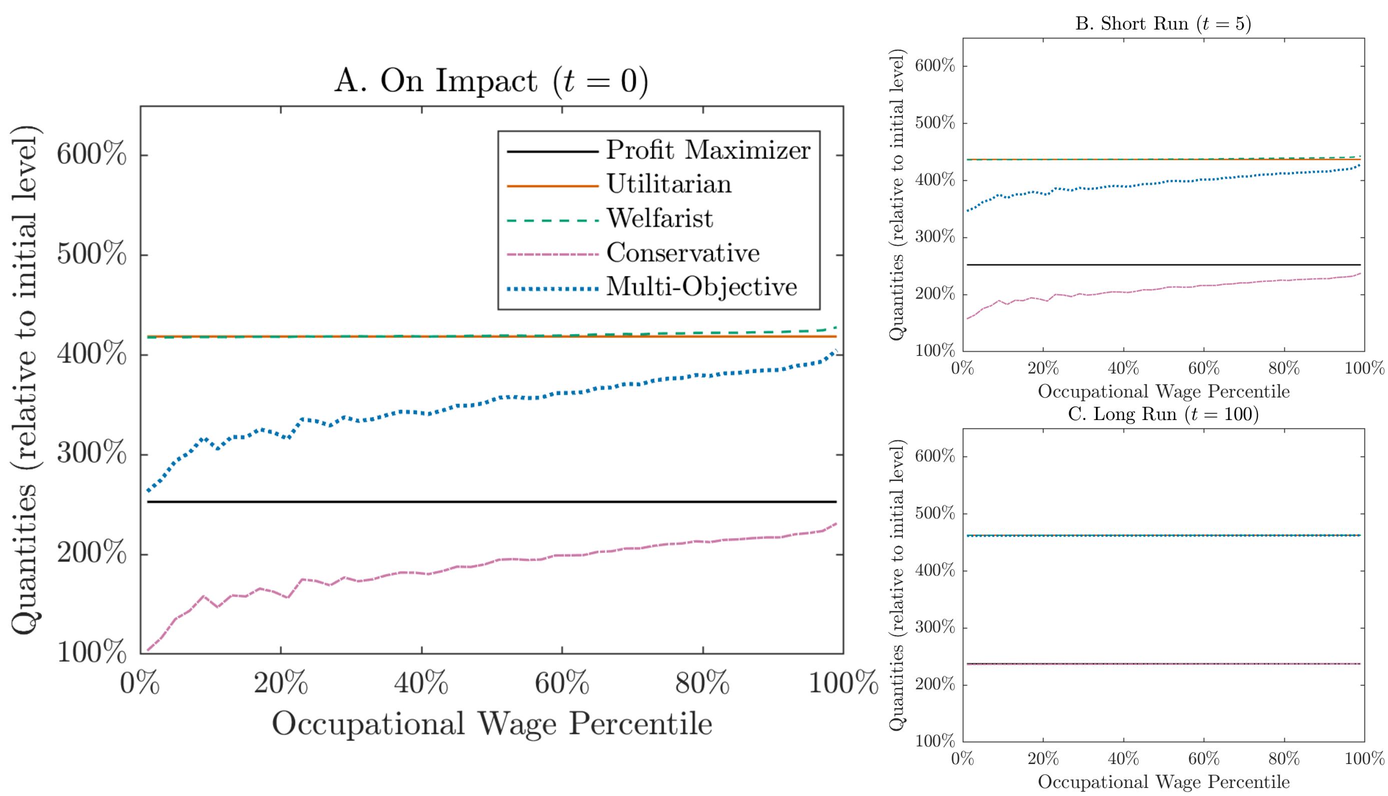

Figure 5: Equilibrium Quantities in the Short and Long Run

Notes: Panel A reports equilibrium quantities on impact ($t = 0$) for AIs capable of automating different occupations (in the horizontal axis) as the percent deviation from pre-AI production levels. The curves are smoothed by binning occupations into 50 quantiles and reporting the average within each bin. Each panel shows five curves, one for each type of firm. Panels B and C report the same curves after 5 and 100 years.

2.2 Why do distributional considerations play such a small role?

Why do distributional considerations play such a small role, especially for low-pay jobs? Two forces explain this finding.

First, and as discussed in Proposition 2, the strength of distributional motives vanishes as AI use deepens. As shown in Figure 5, AI output for the utilitarian and welfarist firms is already 4 times that supplied by workers at baseline. This pushes the firm to prioritize aggregate efficiency over distributional considerations.

Second, the distributional effects of changing wages in a given occupation are not as large as one may have thought, especially at the bottom. To illustrate this point, let's compute the distributional gains of increasing income in occupation s, given by the normalized Pareto weights of the average employee:

$$\text{Average Pareto Weight}_s = \sum_h g_h \frac{\bar{n}_s^h}{\bar{n}_s}.$$

Panel A in Figure 6 reports these average weights for the 525 detailed occupations in our data. Panel B complements this information by plotting the average income percentile of people employed in each occupation.

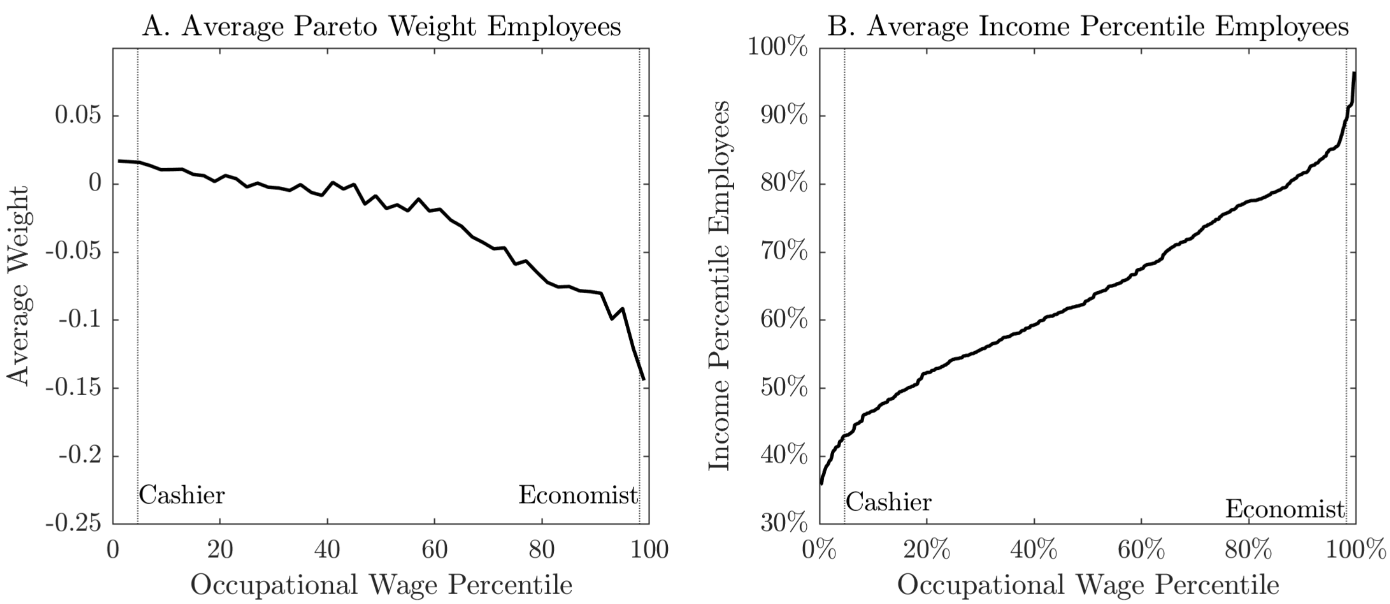

Figure 6: Welfare Weights Across Occupations

Notes: Panel A plots the distributional gains of increasing income by \$1 across occupations. These are computed as $\sum_h g_h \frac{\bar{n}_s^h}{\bar{n}_s}$, where the normalized Pareto weights g^h are from Lockwood and Weinzierl (2016). Panel B plots the average income percentile of workers within an occupation.

All occupations at the bottom half of the pay distribution have positive average weights, showing that increasing income in these jobs has a positive distributional benefit. However, the average weights are small and close to zero, suggesting these benefits are small in practice. Increasing income in jobs at the bottom carries a tiny distributional gain of 1 cent for every dollar. To understand why, consider *Cashiers*—one of the 5% lowest paying jobs in the US. Despite its low pay, people working as cashiers come from households with a wide range of incomes, spanning from the very bottom to the 70th percentile. On average, people employed as cashiers come from households at the 42nd percentile of the income distribution. This lack of segmentation at the bottom implies that protecting cashiers and other low-wage jobs benefits a wide range of households, not just the very poor. This effect is further compounded by the fact that many of the poorest households earn no labor income at all, and are therefore not exposed to the substituting effects from AI.

On the other hand, increasing income in jobs at the top carries a more sizable distributional penalty of 12 cents for every dollar, reflecting the higher degree of income segmentation at the top. Consider *Economists*, one of the 5% highest paid jobs. People in this field typically hail from households at the upper end of the income distribution, with the average economist located at the 90th percentile. This asymmetry explains why distributional concerns matter very little at the bottom but have a more appreciable (though still modest) effect for AI pricing at the top.

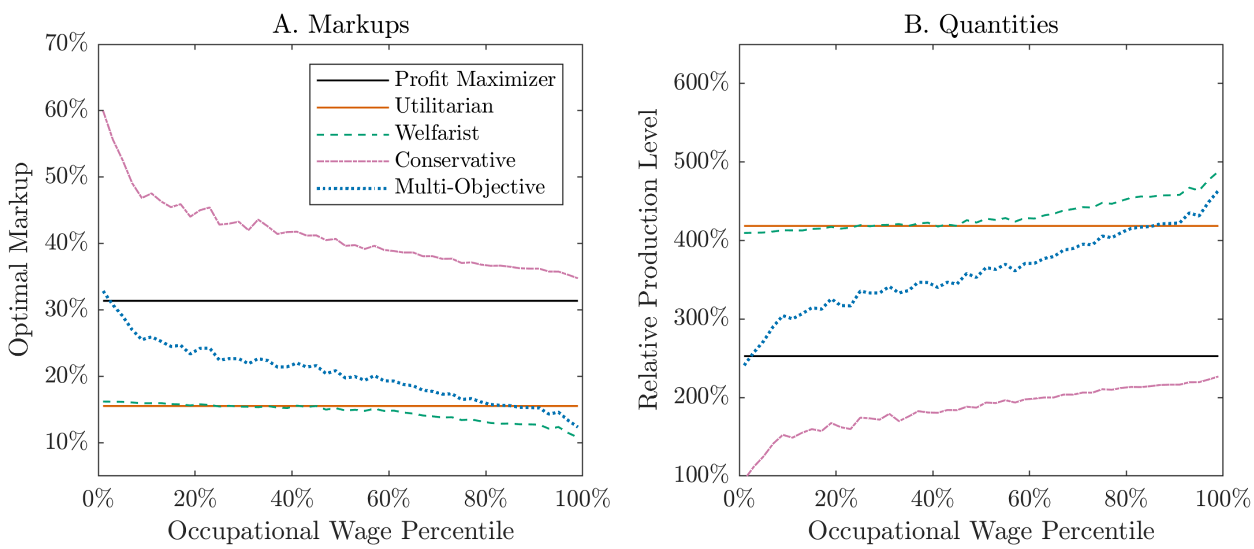

Figure 7: Markups and Quantities on Impact for Stronger Redistributive Preferences

Notes: This figure reports optimal markups and quantities on impact ($t = 0$) for stronger redistributive preferences than baseline, $\tilde{g}_h = 10 \times g_h$. Panel A reports optimal markups and Panel B the associated quantities for the automated skill.

Would distributional considerations for jobs at the bottom matter if the AI firm were even more progressive? Imagine a firm whose welfare weights g^h are ten times those estimated by Lockwood and Weinzierl (2016) in Figure 2. This hypothetical firm is in effect ten times more progressive than the US political system. The resulting equilibrium markups are shown in Figure 7. The stronger distributional considerations call for one percentage point higher markups on AIs that substitute for bottom occupations, relative to what a utilitarian firm would do. However, the incentive to increase aggregate efficiency remains dominant, and it is still optimal for a welfarist firm to expand the quantity of AI produced to broaden access, despite its potential adverse distributional effects at the lower end.

Stronger distributional considerations do make a difference for jobs at the top. A welfarist firm with ten times stronger distributional concerns should charge markups that are half of what a utilitarian firm would charge and produce 20% more output. This is because AIs that substitute for jobs at the top redistribute from workers at the very top (who tend to hold highly paid jobs) towards the rest of the population (who are not exposed to these top jobs)—an extremely valuable proposition from the firm's viewpoint.

In summary, stronger distributional concerns lead to a more aggressive deployment of AI at the top, but have no significant implications for AIs that substitute for jobs at the bottom.

2.3 Should more productive AI be priced differently?

Our baseline results considered the optimal deployment of AIs capable of replacing workers at 50% of their cost. Suppose that $\psi_s = .2\ \bar{w}_s$, so that AI can replace workers at 20% of their cost. How should these more productive AIs be priced and deployed?

Figure 8 reports equilibrium prices and quantities for such AIs across occupations at $t = 0$. Relative to our baseline, markups are slightly higher, while quantities are an order of magnitude larger. This is because a more productive AI firm experiences less competition from workers and captures a greater share of the market, allowing it to charge higher markups.

More importantly, the figure shows that both distributional and labor-market stability motives are weaker when pricing more productive AIs. This can be seen from the fact that outcomes for conservative firms are close to those of a pure profit maximizer, and outcomes for the welfarist and multi-objective firms are close to the utilitarian one. As discussed in Proposition 2, this is because profit and efficiency motives scale with the quantity of AI used, while distributional and stability

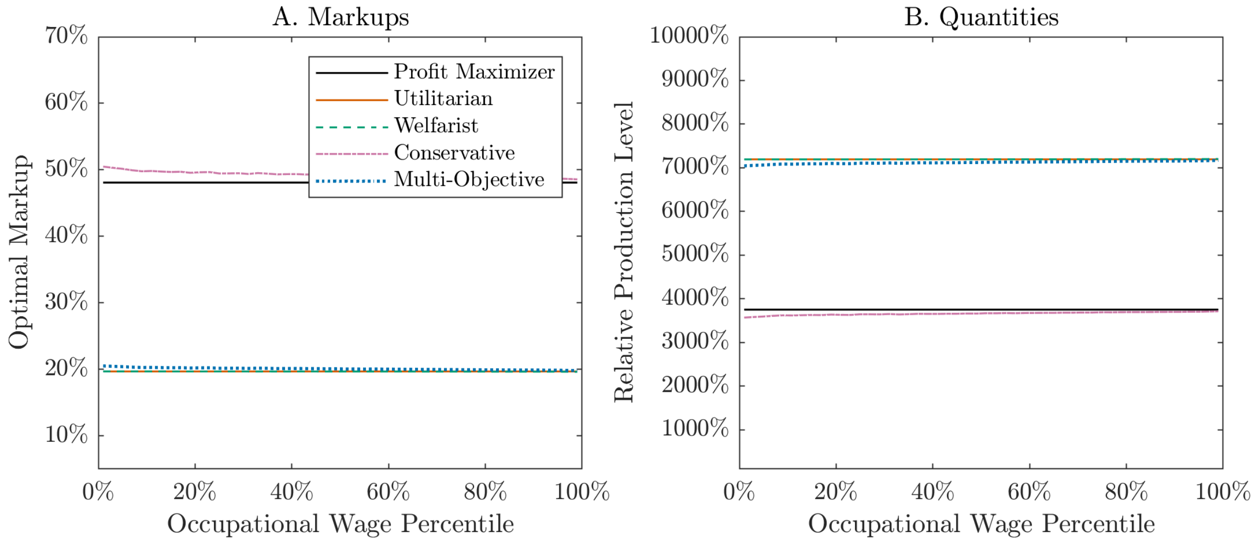

Figure 8: Markups and Quantities on Impact for Highly Productive AI

Notes: This figure reports optimal markups and quantities on impact ($t = 0$) for more productivity AI than baseline, $\psi_s = 0.2$. Panel A reports optimal markups and Panel B the associated quantities for the automated skill.

motives do not.

We conclude that as firms develop more productive and less costly AIs, distributional or stability considerations become less pressing. A socially minded firm with a sufficiently productive AI should behave essentially as a utilitarian one, balancing access and profits only.

2.4 What would a planner do?

To conclude our empirical exploration, we contrast the optimal deployment plan pursued by a socially minded firm with that of a social planner that can control the supply of AIs but has no other tools. Suppose the objective of the planner is to maximize social welfare and maintain labor market stability, giving a weight μ to the utility of financiers—the same given to the average household. The same derivations we did for an AI firm above imply that the planner supplies AI until

$$1 - \frac{\psi_{st}}{w_{st}} = \left(\int_h g^h \frac{\bar{n}_s^h}{q_{st}} e^{-\alpha t} \, dh + \frac{\lambda}{\mu} \int_h \frac{1}{\bar{w}^h} \frac{\bar{n}_s^h}{q_{st}} e^{-\alpha t} \, dh \right) \frac{q_{st}}{q_{st} + \bar{n}_s e^{-\alpha t}} \frac{1}{\sigma_s} \tag{6}$$

This equation states that the planner trades off the reduction in aggregate efficiency from reducing output below its competitive level (the left side) with the distributional and stability gains this creates (the right side).

Figure 9 depicts the optimal deployment plan that a social planner controlling the supply of AI would choose. To ease the comparison with our previous findings, we report the implied markup that would decentralize the planner's allocation. These can also be interpreted as the optimal tax that such a planner would levy on AIs substituting for jobs at the bottom and top of the income distribution, respectively. The figure also reports the markups that a socially minded multi-objective firm would charge (using the baseline values for μ and λ from above).

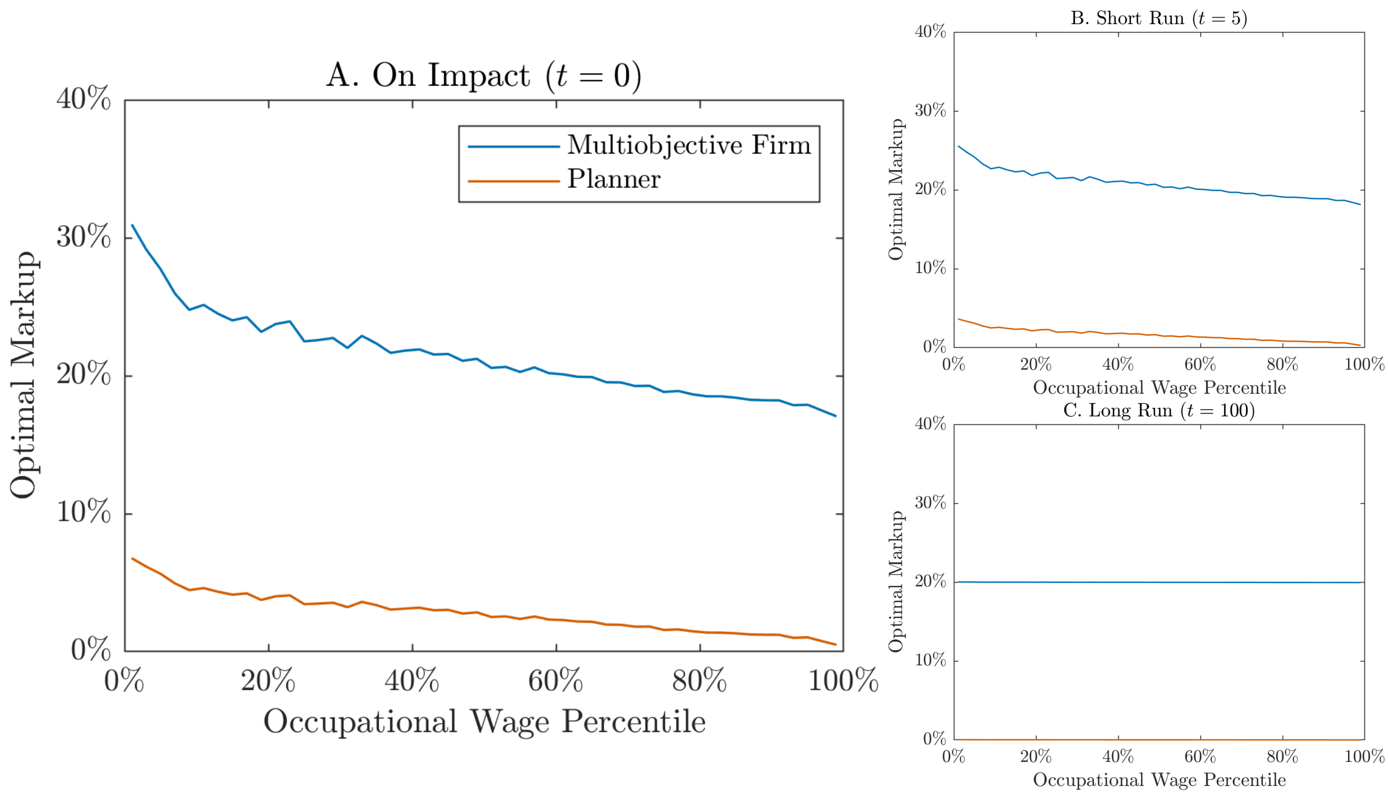

Figure 9: Implied Markups or AI Taxes that Decentralizing the Planner's Allocation

Notes: The figure reports the optimal markup that a social planner would set to balance aggregate efficiency with distributional and stability concerns. Panel A reports markups on impact ($t = 0$). Panels B and C report the same curves after 5 and 100 years. For comparison, the figure also depicts the deployment path followed by a multi-objective firm with $\mu = \lambda = 0.5$.

A conservative social planner that cares about welfare and stability equally ($\mu = \lambda$ as our multi-objective firm), would set an optimal markup on AIs at the bottom of 7% and at the top of less than 1% on impact. Over time, as distributional and stability concerns subside, the conservative social planner would impose no markups or taxes on AI and would implement a competitive outcome, whereas a responsible-AI firm would continue to charge a positive markup, as it balances broadening access with its private profit incentives.

The main result here is that the planner allocation features lower AI prices than what a responsible-AI firm with the same social objectives would charge. Despite its social inclinations, socially responsible firms remain constrained by their private profit motives in how much they can lower prices to broaden access.

2.5 Do our conclusions apply to occupations with high AI replacement risk?

Our results characterize the optimal deployment path for AIs capable of substituting for labor across various occupations, from cashiers to economists, without taking a stance on which jobs could be automated first. Do our conclusions apply to occupations with the highest risk of automation by AI, as identified in existing prospective analyses?

Figure 10: Distributional and Stability Considerations for Occupations at Risk of Replacement

Notes: Panel A plots the redistributive motive, $\mu \ \sum_h g_h \ \frac{\bar{n}_s^h}{\bar{n}_s}$, against the share of tasks automatable by AI following Eloundou et al. (2023). Panel B plots the non-disruption motive, $\lambda \ \sum_h \frac{1}{\bar{w}_h} \ \frac{\bar{n}_s^h}{\bar{n}_s}$ against the same AI measure.

We answer this question using data from Eloundou et al. (2023) on the share of core tasks by occupation that could be automated with LLM-powered systems. Figure 10 shows that occupations at risk (in the horizontal axis) do not stand out in their distributional or stability considerations. Panel A shows that highly exposed occupations have average Pareto weights near zero. Panel B shows that the average cost of disruptions among employees does not systematically vary with AI exposure either. This is because prospective studies, such as Eloundou et al. (2023), suggest that the set of occupations at risk is spread throughout the income distribution and does not concentrate at either the bottom or the top.

In sum, the conclusions drawn above for the entire universe of jobs apply equally well to the subset of occupations more at risk of being substituted by AIs.

3 Theory Extensions

This section explores theoretical extensions. First, we discuss how taxes and the safety net affect the deployment of AI by socially minded firms. Second, we discuss the possibility that some AIs may not replicate human skill but could eventually acquire new capabilities that allow these systems to produce entirely new goods and services without devaluing existing human skills. Finally, we discuss the case when multiple AI firms are competing a la Cournot. The proofs for this extension are in the appendix.

3.1 Taxes and the safety net

We now extend the model to account for the tax system and the safety net. Assume the after-tax labor income of household h is

$$\text{After-tax labor income}_t^h \equiv \mathcal{T}(w_t^h) + T_t,$$

where T_t is a common transfer that balances the government budget and $\mathcal{T}(.)$ is an increasing tax function, with $\mathcal{T}(0) = 0$ and $1 - \mathcal{T}'(w_t^h) > 0$ giving the marginal tax rate experienced by households at different points of the income distribution.

The AI firm's objective function is now

$$V = \text{PDV } \pi_t \ + \int_h \mu^h \ u^h \ dh \ + \ \lambda \int_{h: w_t^h < \bar{w}^h} \text{PDV } \frac{\mathcal{T}(w_t^h)}{\mathcal{T}(\bar{w}^h)} \, dh,$$

where we assume that the stability term depends on how actions by the AI firm reduce after-tax labor income $\mathcal{T}(w_t^h)$ relative to its status quo level $\mathcal{T}(\bar{w}^h)$.

Proposition 3. *In the quasi-linear case with government taxes, a socially-responsible firm produces q_{st} until*

$$\mathcal{L}_{st} = \Big(1 - \mu \ + \ \mu \int_h g^h \, \mathcal{T}'(w_t^h) \, \frac{n_{st}^h}{q_{st}} \, dh \ + \ \lambda \int_h \frac{\mathcal{T}'(w_t^h)}{\mathcal{T}(\bar{w}^h)} \, \frac{n_{st}^h}{q_{st}} \, dh \Big) \, \frac{1}{\varepsilon_{st}} \tag{7}$$

The proposition shows that a more progressive tax system (as evidenced by a *lower* keep rate $\mathcal{T}'(w_t^h)$) weakens the firm's distributional and stability concerns. In the extreme case of full redistribution (i.e., $\mathcal{T}'(w_t^h) = 0$), distributional and stability considerations vanish. This highlights an important interplay between public policy and self-regulation. The more we redistribute via the tax system, the less an AI firm should worry about its downstream distributional effects and the more it should prioritize broadening access.

3.2 AI as creating new goods and services

Our formulas assume that AIs substitute for human labor in existing jobs, a natural application since these systems are trained on human-generated data to mimic us. Yet some argue that large models, when trained on vast datasets, can develop novel capabilities and produce goods and services that surpass anything humans have created so far.

To account for this possibility, assume the firm also develops AIs that create new goods and services, such as new proteins that it can sell or license to medical laboratories. Assume also that household utility is given by

$$u^h(c) = c_{0t}^h + \sum_{s \in \mathcal{S}} \gamma_s^{1/\sigma_s} \, \frac{c_s^{1-1/\sigma_s}}{1 - 1/\sigma_s} + \sum_{s' \in \mathcal{N}} \gamma_{s'}^{1/\sigma_{s'}} \, \frac{c_{s'}^{1-1/\sigma_{s'}}}{1 - 1/\sigma_{s'}},$$

for $\sigma_{s'} > 1$. The set $\mathcal{N}$ represents new goods and services produced by AIs, indexed by s'. We let $n_{s'}^h = 0$ for all $s' \in \mathcal{N}$, indicating that humans were not able to produce these novel goods and services.

Proposition 4. *In an economy where (a)–(d) hold, the optimal pricing of novel AIs $s' \in \mathcal{N}$ satisfies a modified Lerner rule*

$$\mathcal{L}_{s't} = \Big(1 \ - \ \mu\Big) \, \frac{1}{\sigma_{s'}}. \tag{8}$$

AIs that expand the range of goods and services benefit all workers without disrupting existing labor markets, and thus raise no concerns regarding distribution or stability. For this class of AIs, the main responsibility of a socially minded firm is to price close to marginal cost and broaden access.

3.3 Competition among AI producers

We extend the baseline model in Section 1.2 to incorporate competition among AI companies. For each $s \in \mathcal{A}$, suppose $M_s > 1$ identical firms produce the AI and compete in quantities à la Cournot.

Proposition 5. *In an economy where (a)–(d) hold and M_s symmetric companies compete in quantities, the equilibrium price of AI satisfies*

$$\mathcal{L}_{st} = \left(\frac{1 - \mu}{M_s} + \mu \int_h g^h \, \frac{\bar{n}_s^h}{q_{st}} \, e^{-\alpha t} \, dh + \lambda \int_h \frac{1}{\bar{w}^h} \, \frac{\bar{n}_s^h}{q_{st}} \, e^{-\alpha t} \, dh \right) \frac{q_{st}}{q_{st} + \bar{n}_s \, e^{-\alpha t}} \, \frac{1}{\sigma_s}, \tag{9}$$

where q_{st} is the aggregate quantity of AI used in $s \in \mathcal{A}$.

As usual, competition forces firms to set prices closer to their marginal cost and expand quantities. This is evidenced by the fact that, in (9), the term "$1 - \mu$" is divided by M_{st}.

The proposition also shows that distributional and stability concerns become increasingly relevant as competition between AI suppliers intensifies. To see this, note that a pure-profit-maximizing firm would price according to

$$\mathcal{L}_{st} = \frac{1}{M_{st}} \, \frac{q_{st}}{q_{st} + \bar{n}_s \, e^{-\alpha t}} \, \frac{1}{\sigma_s}$$

The more competition this firm faces, the closer it would price to marginal cost. Consider now a socially minded firm. As competition intensifies, distributional and stability considerations become the sole forces causing the firm to deviate from profit-maximizing pricing. Broadening access is no longer a first-order concern.

The reason why this happens is that competition pushes firms to produce closer to the efficient level of AI (from an aggregate efficiency point of view). Starting from this level, the aggregate efficiency gains from further expanding access are limited, since firms are already pricing close to their marginal cost. Instead, the incentive to curb quantities to limit adverse distributional effects or maintain stability remains active and becomes the dominant force guiding firms' actions.

4 Conclusion

How should socially minded firms deploy and price their AIs? This paper provides a framework to address this question by extending Lerner's Rule to incorporate a broader set of objectives: generating profits, promoting social welfare, and minimizing labor-market disruptions. The resulting pricing formulas clarify how these objectives shape markups, deployment speed, and access to AI.

Applying our framework to US data across hundreds of occupations, we find that a firm that cares about welfare and stability equally should price near the profit-maximizing level in the short run, but closer to marginal cost over time. This gradual strategy limits short-run disruptions and balances them with improved access to the technology in the medium run.

Our conclusion from this exercise is that the most pro-social course of action for a monopolist AI firm is to refrain from exploiting its market power, except in the very short run, when stability considerations are the most pressing. This conclusion is particularly relevant for AIs that produce new goods and services without disrupting existing labor markets, and when the government is already engaged in effective redistribution.

This conclusion also depends on the baseline level of competition in AI markets. Whether the incentive to prioritize access dominates depends on how constrained supply was by the exercise of market power to begin with. If AI firms face little competition, the incentive to broaden access by lowering prices dominates the actions of a socially responsible firm. Instead, if competition among AI firms already results in quantities that are close to efficient, the incentive to broaden access loses relevance, while distributional and stability concerns become dominant.

Because our analysis focuses solely on labor market and economic efficiency considerations, abstracting from broader societal and existential risks, our conclusions apply only to well-aligned or narrow AIs without existential risks.

References

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

Acemoglu, Daron, Andrea Manera, and Pascual Restrepo. 2020. "Does the US Tax Code Favor Automation?" *Brookings Papers on Economic Activity*, 231–285.

Acemoglu, Daron, and Todd Lensman. 2024. "Regulating Transformative Technologies." *American Economic Review: Insights*, 6(3): 359–76.

Beraja, Martin, and Nathan Zorzi. 2022. "Inefficient Automation." National Bureau of Economic Research Working Paper 30154.

Bils, Mark, Bariş Kaymak, and Kai-Jie Wu. 2024. "Labor Substitutability among Schooling Groups." *American Economic Journal: Macroeconomics*, 16(4): 1–34.

Bond, Philip, and Lukas Kremens. 2025. "Income and inequality under asymptotically full automation."

Broda, Christian, and David E. Weinstein. 2006. "Globalization and the Gains From Variety." *The Quarterly Journal of Economics*, 121(2): 541–585.

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond. 2023. "Generative AI at Work."

Brynjolfsson, Erik, Tom Mitchell, and Daniel Rock. 2018. "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?" *AEA Papers and Proceedings*, 108: 43–47.

Burstein, Ariel, Eduardo Morales, and Jonathan Vogel. 2019. "Changes in Between-group Inequality: Computers, Occupations, and International Trade." *American Economic Journal: Macroeconomics*, 11(2): 348–400.

Corden, W. Max. 1974. *Trade Policy and Economic Welfare.* Oxford: Clarendon Press.

Costinot, Arnaud, and Iván Werning. 2022. "Robots, Trade and Luddism: A Sufficient Statistic Approach to Optimal Technology Regulation." *The Review of Economic Studies*.

Donald, Eric. 2022. "Optimal Taxation with Automation: Navigating Capital and Labor's Complicated Relationship." Boston University Mimeo.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models."

Felten, Ed, Manav Raj, and Robert Seamans. 2023. “How will Language Modelers like ChatGPT Affect Occupations and Industries?”

Felten, Edward, Manav Raj, and Robert Seamans. 2021. “Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses.” *Strategic Management Journal*, 42: 2195–2217.

Friedman, Milton. 1970. “A Friedman doctrine– The Social Responsibility of Business Is to Increase Its Profits.” *The New York Times*.

Guerreiro, Joao, Sergio Rebelo, and Pedro Teles. 2021. “Should Robots Be Taxed?” *The Review of Economic Studies*, 89(1): 279–311.

Guerreiro, Joao, Sergio Rebelo, and Pedro Teles. 2023. “Regulating Artificial Intelligence.” National Bureau of Economic Research Working Paper 31921.

Handa, Kunal, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. 2025. “Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations.”

Jones, Charles I. 2024. “The AI Dilemma: Growth versus Existential Risk.” *American Economic Review: Insights*, 6(4): 575–90.

Jones, Charles I. 2025. “How Much Should We Spend to Reduce A.I.’s Existential Risk?” National Bureau of Economic Research Working Paper 33602.

Katz, Lawrence F, and Kevin M Murphy. 1992. “Changes in Relative Wages, 1963–1987: Supply and Demand factors.” *The Quarterly Journal of Economics*, 107(1): 35–78.

Lehr, Nils H, and Pascual Restrepo. 2024. “Optimal Gradualism.” National Bureau of Economic Research Working Paper 30755.

Lockwood, Benjamin B., and Matthew Weinzierl. 2016. “Positive and normative judgments implicit in U.S. tax policy, and the costs of unequal growth and recessions.” *Journal of Monetary Economics*, 77: 30–47. “Inequality, Institutions, and Redistribution”held at the Stern School of Business,New York University,April 24-25, 2015.

Moll, Benjamin, Lukasz Rachel, and Pascual Restrepo. 2022. "Uneven Growth: Automation's Impact on Income and Wealth Inequality." *Econometrica*, 90(6): 2645–2683.

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

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot."

Thuemmel, Uwe. 2023. "Optimal Taxation of Robots." *Journal of the European Economic Association*, 1(3): 1154–1190.

Webb, Michael. 2020. "The Impact of Artificial Intelligence on the Labor Market."

Appendix

This appendix derives equation (1) and also the planner solution in (6). It then provides proofs for the extensions.

Derivation of (1): First, note that $r_t = \rho$, since financiers must be indifferent between consuming or saving.

Following an arbitrary change in quantities by the AI firm, we get

$$\delta V = \int_0^t e^{-\rho t} \Bigg\{ \sum_{s \in \mathcal{A}} \left(q_{st}\ \delta w_{st} + \left(w_{st} - \psi_{st} \right) \delta q_{st} \right) \\ + \int_h \mu^h \Big(\sum_s n_{st}^h\ \delta w_{st} - \sum_{j \neq 0} c_{jt}^h\ \delta p_{jt} \Big)\, dh \\ + \lambda \int_{h: w_t^h < \bar{w}^h} \frac{1}{\bar{w}^h} \sum_s n_{st}^h\ \delta w_{st}\ dh \Bigg\}\, dt.$$

In this expression:

- The first line indicates the changes in profits that flow to financiers.
- The second line gives the change in households' utility. By assumption, $c_{0t}^h > 0$, which implies that the marginal value of income in period t is $e^{-\rho t}$. This is then multiplied by μ^h, capturing the social value of increasing utility for household h, and the change in household net income resulting from the perturbation. Note that while households adjust their consumption and savings decisions in response to price changes, these changes are second-order due to the envelope theorem. This is why only the change in income resulting from price changes is reflected. Note also that commodity $j = 0$ is the numeraire, and so its price is fixed at one.
- The third line gives the effects via labor market disruptions, which are assumed to be a function of wages. Note that this applies to all households, as AI necessarily reduces nominal wages for all households with $n_{st}^h > 0$.

We can rewrite the above expression as

$$\delta V = \int_0^t e^{-\rho t} \Big\{ (1-\mu) \sum_{s\in\mathcal{A}} \big(q_{st}\ \delta w_{st} + \big(w_{st} - \psi_{st} \big)\ \delta q_{st} \big) \ + \ \mu \sum_{s\in\mathcal{A}} \big(w_{st} - \psi_{st} \big)\ \delta q_{st}$$

$$+ \ \mu \left(\sum_{s\in\mathcal{A}} q_{st}\ \delta w_{st} \ + \ \int_h \big(\sum_s n^h_{st}\ \delta w_{st} - \sum_{j\neq 0} c^h_{jt}\ \delta p_{jt} \big)\ dh \right)$$

$$+ \ \mu \int_h g^h\ \big(\sum_s n^h_{st}\ \delta w_{st} - \sum_{j\neq 0} c^h_j\ \delta p_{jt} \big)\ dh$$

$$+ \ \lambda \int_{h: w^h_t < \bar{w}^h} \frac{1}{\bar{w}^h}\ \sum_s n^h_{st}\ \delta w_{st}\ dh \Big\}\ dt.$$

Using $\ell_{st} = \int_h n^h_{st}\ dh \ + q_{st}$ for $s \in \mathcal{A}$ and $\ell_{st} = \int_h n^h_{st}\ dh$ for $s \notin \mathcal{A}$), plus market clearing for commodities $j \neq 0$, this simplifies to:

$$\delta V = \int_0^t e^{-\rho t} \Big\{ (1-\mu) \sum_{s\in\mathcal{A}} \big(q_{st}\ \delta w_{st} + \big(w_{st} - \psi_{st} \big)\ \delta q_{st} \big) \ + \ \mu \sum_{s\in\mathcal{A}} \big(w_{st} - \psi_{st} \big)\ \delta q_{st}$$

$$+ \ \mu \left(\sum_s \ell_{st}\ \delta w_{st} \ - \sum_{j\neq 0} y_{jt}\ \delta p_{jt} \right)$$

$$+ \ \mu \int_h g^h\ \big(\sum_s n^h_{st}\ \delta w_{st} - \sum_{j\neq 0} c^h_{jt}\ \delta p_{jt} \big)\ dh$$

$$+ \ \lambda \int_{h: w^h_t < \bar{w}^h} \frac{1}{\bar{w}^h}\ \sum_s n^h_{st}\ \delta w_{st}\ dh \Big\}\ dt.$$

Because the production of commodities is competitive and features constant returns to scale, firms make zero profits, and the envelope theorem (applied to their profits) implies

$$\sum_{j\neq 0} y_{jt}\ \delta p_{jt} - \sum_s \ell_{st}\ \delta w_{st} = 0.$$

That is, the second line in the equation for δV is zero. To conclude, note that $c^h_{jt} = y_{jt}$ for $j \neq 0$ in the third line because of quasi-linearity. The term $\sum_{j\neq 0} c^h_{jt}\ \delta p_{jt}$ is then common to all households and cancels because $\int_h g^h\ dh = 0$. This simplification yields (1).

Derivation of (6): Let

$$S = \mu\, u \;+\; \int_h \mu^h\, u^h\, dh \;+\; \lambda \int_{h:w_t^h<\bar{w}^h} \text{PDV}\, \frac{w_t^h}{\bar{w}^h}\, dh$$

be the planners' objective. Consider a small perturbation in the quantity of AI produced. This affects the equilibrium value of S as follows:

$$\begin{aligned}\delta S = \int_0^t e^{-\rho t} \Big\{ \mu \sum_{s\in\mathcal{A}} \left(q_{st}\, \delta w_{st} + \left(w_{st} - \psi_{st}\right) \delta q_{st}\right) \\ + \int_h \mu^h \,\Big(\sum_s n_{st}^h\, \delta w_{st} - \sum_{j\neq 0} c_{jt}^h\, \delta p_{jt}\Big)\, dh \\ + \lambda \int_{h:w_t^h<\bar{w}^h} \frac{1}{\bar{w}^h}\, \sum_s n_{st}^h\, \delta w_{st}\, dh \Big\}\, dt.\end{aligned}$$

We can rewrite the above expression as

$$\begin{aligned}\delta S = \int_0^t e^{-\rho t} \Big\{ \mu \Big(\sum_{s\in\mathcal{A}} \left(q_{st}\, \delta w_{st} + \left(w_{st} - \psi_{st}\right) \delta q_{st}\right) \;+\; \int_h \Big(\sum_s n_{st}^h\, \delta w_{st} - \sum_{j\neq 0} c_{jt}^h\, \delta p_{jt}\Big)\, dh \Big) \\ + \mu \int_h g^h \,\Big(\sum_s n_{st}^h\, \delta w_{st} - \sum_{j\neq 0} c_j^h\, \delta p_{jt}\Big)\, dh \\ + \lambda \int_{h:w_t^h<\bar{w}^h} \frac{1}{\bar{w}^h}\, \sum_s n_{st}^h\, \delta w_{st}\, dh \Big\}\, dt.\end{aligned}$$

Using $\ell_{st} = \int_h n_{st}^h\, dh \; + q_{st}$ for $s \in \mathcal{A}$ and $\ell_{st} = \int_h n_{st}^h\, dh$ for $s \notin \mathcal{A}$), plus market clearing for commodities $j \neq 0$, the first line simplifies to:

$$\begin{aligned}\delta S = \int_0^t e^{-\rho t} \Big\{ \mu \Big(\sum_{s\in\mathcal{A}} \left(w_{st} - \psi_{st}\right) \delta q_{st} \;+ \sum_s \ell_{st}\, \delta w_{st} \;- \sum_{j\neq 0} y_{jt}\, \delta p_{jt} \Big) \\ + \mu \int_h g^h \,\Big(\sum_s n_{st}^h\, \delta w_{st} - \sum_{j\neq 0} c_{jt}^h\, \delta p_{jt}\Big)\, dh \\ + \lambda \int_{h:w_t^h<\bar{w}^h} \frac{1}{\bar{w}^h}\, \sum_s n_{st}^h\, \delta w_{st}\, dh \Big\}\, dt.\end{aligned}$$

Because the production of commodities is competitive and features constant returns to scale, firms make zero profits, and the envelope theorem (applied to their profits) implies

$$\sum_{j\neq 0} y_{jt}\ \delta p_{jt} - \sum_{s} \ell_{st}\ \delta w_{st} = 0.$$

Moreover, $c^h_{jt} = y_{jt}$ for $j \neq 0$ because of quasi-linearity. The term $\sum_{j\neq 0} c^h_{jt}\ \delta p_{jt}$ is then common to all households and cancels because $\int_h g^h\ dh = 0$. Making both replacements in the equation for δS yields the planner's variant of (1):

$$\delta S = \int_0^t e^{-\rho t} \Big\{ \mu \sum_{s\in\mathcal{A}} \left(w_{st} - \psi_{st}\right) \delta q_{st}\ +\ \mu \int_h g^h\ \Big(\sum_s n^h_{st}\ \delta w_{st} - \sum_{j\neq 0} c^h_{jt}\ \delta p_{jt}\Big)\ dh$$
$$+\ \lambda \int_{h:w^h_t<\bar{w}^h} \frac{1}{\bar{w}^h}\ \sum_s n^h_{st}\ \delta w_{st}\ dh \Big\}\ dt.$$

At an optimum, the planner sets quantities so that $\delta S = 0$. This implies that for all t and $s \in \mathcal{A}$,

$$0 = \mu\ \left(w_{st} - \psi_{st}\right) \delta q_{st}\ +\ \mu \int_h g^h\ \sum_{s'} n^h_{s't}\ \delta w_{s't}\ dh\ +\ \lambda \int_{h:w^h_t<\bar{w}^h} \frac{1}{\bar{w}^h}\ \sum_{s'} n^h_{s't}\ \delta w_{s't}\ dh,$$

which in the simplified economy considered in Proposition 2 can be written as in equation (6).

Proof of Proposition 3: Following an arbitrary change in quantities by the AI firm, we get

$$\delta V = \int_0^t e^{-\rho t} \Big\{ \sum_{s\in\mathcal{A}} \left(q_{st}\ \delta w_{st} + \left(w_{st} - \psi_{st}\right) \delta q_{st}\right)$$
$$+\ \int_h \mu^h\ \Big(\mathcal{T}'(w^h_t)\ \sum_s n^h_{st}\ \delta w_{st} + \delta T_t - \sum_{j\neq 0} c^h_{jt}\ \delta p_{jt}\Big)\ dh$$
$$+\ \lambda \int_{h:w^h_t<\bar{w}^h} \frac{\mathcal{T}'(w^h_t)}{\mathcal{T}(\bar{w}^h)}\ \sum_s n^h_{st}\ \delta w_{st}\ dh \Big\}\ dt.$$

Using the fact that

$$\delta T_t\ =\ \int_h \Big(\sum_s n^h_{st}\ \delta w_{st} - \mathcal{T}'(w^h_t)\ \sum_s n^h_{st}\ \delta w_{st}\Big)\ dh,$$

which follows from the requirement that tax revenue is rebated to households, the above expression for δV can be rewritten as

$$\delta V = \int_0^t e^{-\rho t} \Bigg\{ (1-\mu) \sum_{s\in\mathcal{A}} \left(q_{st}\ \delta w_{st} + \left(w_{st} - \psi_{st}\right) \delta q_{st} \right) \ +\ \mu \sum_{s\in\mathcal{A}} \left(w_{st} - \psi_{st}\right) \delta q_{st}$$

$$+\ \mu \left(\sum_{s\in\mathcal{A}} q_{st}\ \delta w_{st}\ +\ \int_h \Big(\sum_s n_{st}^h\ \delta w_{st} - \sum_{j\neq 0} c_{jt}^h\ \delta p_{jt} \Big)\ dh \right)$$

$$+\ \mu \int_h g^h\ \Big(\mathcal{T}'(w_t^h) \sum_s n_{st}^h\ \delta w_{st} - \sum_{j\neq 0} c_j^h\ \delta p_{jt}\Big)\ dh$$

$$+\ \lambda \int_{h:w_t^h<\bar{w}^h} \frac{\mathcal{T}'(w_t^h)}{\mathcal{T}(\bar{w}^h)} \sum_s n_{st}^h\ \delta w_{st}\ dh \Bigg\}\ dt.$$

From here on, we follow the same steps from the derivation of equation (1) to obtain the variant:

$$\delta V = \int_0^t e^{-\rho t} \Bigg\{ (1-\mu) \sum_{s\in\mathcal{A}} \left(q_{st}\ \delta w_{st} + \left(w_{st} - \psi_{st}\right) \delta q_{st} \right) \ +\ \mu \sum_{s\in\mathcal{A}} \left(w_{st} - \psi_{st}\right) \delta q_{st}$$

$$+\ \mu \int_h g^h\ \mathcal{T}'(w_t^h) \sum_s n_{st}^h\ \delta w_{st}\ dh$$

$$+\ \lambda \int_{h:w_t^h<\bar{w}^h} \frac{\mathcal{T}'(w_t^h)}{\mathcal{T}(\bar{w}^h)} \sum_s n_{st}^h\ \delta w_{st}\ dh \Bigg\}\ dt.$$

From here on, we proceed as in the proof of Proposition 2.

Proof of Proposition 4: Following an arbitrary change in quantities for AIs in $\mathcal{N}$ or $\mathcal{A}$, we get

$$\delta V = \int_0^t e^{-\rho t} \Bigg\{ \sum_{s\in\mathcal{N}\cup\mathcal{A}} \left(q_{st}\ \delta w_{st} + \left(w_{st} - \psi_{st}\right) \delta q_{st} \right)$$

$$+\ \int_h \mu^h\ \Big(\sum_s n_{st}^h\ \delta w_{st} - \sum_{j\neq 0} c_{jt}^h\ \delta p_{jt} \Big)\ dh$$

$$+\ \lambda \int_{h:w_t^h<\bar{w}^h} \frac{1}{\bar{w}^h} \sum_s n_{st}^h\ \delta w_{st}\ dh \Bigg\}\ dt.$$

Following the same steps as in the derivation of equation (1), we can write this as

$$\delta V = \int_0^t e^{-\rho t} \Big\{ (1-\mu) \sum_{s \in \mathcal{N} \cup \mathcal{A}} \big(q_{st}\ \delta w_{st} + (w_{st} - \psi_{st})\ \delta q_{st} \big) \ + \ \mu \sum_{s \in \mathcal{N} \cup \mathcal{A}} (w_{st} - \psi_{st})\ \delta q_{st}$$
$$+ \ \mu \int_h g^h \ \sum_s n_{st}^h \ \delta w_{st}\ dh$$
$$+ \ \lambda \int_{h: w_t^h < \bar{w}^h} \frac{1}{\bar{w}^h} \ \sum_s n_{st}^h \ \delta w_{st}\ dh \Big\}\ dt.$$

At an optimum, the firm sets quantities so that $\delta V = 0$. In an economy with (a)–(d), this implies that for every $s' \in \mathcal{N}$ and time t,

$$0 = (1-\mu) \big(q_{s't}\ \delta w_{s't} + (w_{s't} - \psi_{s't})\ \delta q_{s't} \big) + \ \mu \ (w_{s't} - \psi_{s't})\ \delta q_{s't},$$

where we used the fact that $n_{s't}^h = 0$ and the fact that changing $q_{s't}$ does not affect wages for other skills $s \neq s'$. This expression can then be rearranged into (8).

Proof of Proposition 5: Let $q_{st}^{(i)}$ be the quantity supplied by one of the AI firms in $s \in \mathcal{A}$, where $i = 1, 2, \ldots, M_s$. Denote by $q_{st}^{(-i)}$ the quantity supplied by its competitors. A perturbation in $q_{st}^{(i)}$ changes the firm objective ($V^{(i)}$) by

$$\delta V^{(i)} = \int_0^t e^{-\rho t} \Big\{ \sum_{s \in \mathcal{A}} \big(q_{st}^{(i)}\ \delta w_{st} + (w_{st} - \psi_{st})\ \delta q_{st}^{(i)} \big) + \mu \sum_{s \in \mathcal{A}} q_{st}^{(-i)}\ \delta w_{st}$$
$$+ \int_h \mu^h \ \big(\sum_s n_{st}^h \ \delta w_{st} - \sum_{j \neq 0} c_{jt}^h \ \delta p_{jt} \big)\ dh$$
$$+ \ \lambda \int_{h: w_t^h < \bar{w}^h} \frac{1}{\bar{w}^h} \ \sum_s n_{st}^h \ \delta w_{st}\ dh \Big\}\ dt.$$

This expression assumes the AI firm values the utility of owners of other AI firms at a rate μ, which is the same as the average household.

Rearranging terms, this can be expressed as

$$\begin{aligned}\delta V^{(i)} = \int_0^t e^{-\rho t} \Big\{ &(1-\mu) \sum_{s\in\mathcal{A}} \left(q_{st}^{(i)}\, \delta w_{st} + \left(w_{st} - \psi_{st} \right) \delta q_{st}^{(i)} \right) \; + \; \mu \sum_{s\in\mathcal{A}} \left(w_{st} - \psi_{st} \right) \delta q_{st}^{(i)} \\ &+ \; \mu \left(\sum_{s\in\mathcal{A}} q_{st}\, \delta w_{st} \; + \; \int_h \big(\sum_s n_{st}^h\, \delta w_{st} - \sum_{j\neq 0} c_{jt}^h\, \delta p_{jt} \big)\, dh \right) \\ &+ \mu \int_h g^h\, \big(\sum_s n_{st}^h\, \delta w_{st} - \sum_{j\neq 0} c_{jt}^h\, \delta p_{jt} \big)\, dh \\ &+ \; \lambda \int_{h: w_t^h < \bar{w}^h} \frac{1}{\bar{w}^h} \; \sum_s n_{st}^h\, \delta w_{st}\, dh \Big\}\, dt. \end{aligned}$$

Following the same steps as in the derivation of (1), we have that the second line is zero and the effect of prices on the third line also averages to zero. This yields the variant of (1):

$$\begin{aligned}\delta V^{(i)} = \int_0^t e^{-\rho t} \Big\{ &(1-\mu) \sum_{s\in\mathcal{A}} \left(q_{st}^{(i)}\, \delta w_{st} + \left(w_{st} - \psi_{st} \right) \delta q_{st}^{(i)} \right) \; + \; \mu \sum_{s\in\mathcal{A}} \left(w_{st} - \psi_{st} \right) \delta q_{st}^{(i)} \\ &+ \mu \int_h g^h \; \sum_s n_{st}^h\, \delta w_{st}\, dh \\ &+ \; \lambda \int_{h: w_t^h < \bar{w}^h} \frac{1}{\bar{w}^h} \; \sum_s n_{st}^h\, \delta w_{st}\, dh \Big\}\, dt. \end{aligned}$$

At an optimum, the firm sets quantities so that $\delta V^{(i)} = 0$. In an economy with (a)–(d), this implies that for every $s \in \mathcal{A}$ and time t,

$$0 \; = (1-\mu) \left(q_{st}^{(i)} \frac{\delta w_{st}}{\delta q_{st}^{(i)}} + \left(w_{st} - \psi_{st} \right) \right) \; + \; \mu \left(w_{st} - \psi_{st} \right) + \mu \int_h g^h\, n_{st}^h \frac{\delta w_{st}}{\delta q_{st}^{(i)}}\, dh \; + \; \lambda \int_h \frac{1}{\bar{w}^h}\, n_{st}^h \frac{\delta w_{st}}{\delta q_{st}^{(i)}}\, dh.$$

This can be written as

$$1 - \frac{\psi_{st}}{w_{st}} \; = \; \left(1 \; - \; \mu \; + \; \mu \int_h g^h \frac{n_{st}^h}{q_{st}^{(i)}}\, dh \; + \; \lambda \int_h \frac{1}{\bar{w}^h} \frac{n_{st}^h}{q_{st}^{(i)}}\, dh \right) \frac{1}{\varepsilon_{st}^{(i)}},$$

where $\frac{1}{\varepsilon_{st}^{(i)}} \; = \; -\frac{\partial \ln w_{st}}{\partial \ln q_{st}^{(i)}}$ is the residual elasticity of demand faced by the firm. The iso-elastic specification in Section 1.2 implies

$$\frac{1}{\varepsilon_{st}^{(i)}} = \frac{q_{st}^{(i)}}{q_{st} + \bar{n}_s e^{-\alpha t}} \frac{1}{\sigma_s}.$$

Plugging this expression above and using the fact that $q_{st}^{(i)} = q_{st}/M_s$ (from symmetry), we get (9).