THE PRICE OF INTELLIGENCE: HOW SHOULD SOCIALLY-MINDED FIRMS PRICE AND DEPLOY AI?

By

Nils H. Lehr and Pascual Restrepo

May 2025

COWLES FOUNDATION DISCUSSION PAPER NO. 2445

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

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

Nils H. Lehr
International Monetary Fund

Pascual Restrepo
Yale University

May 20, 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 enhance 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.

Artificial Intelligence (AI) is transforming the economy, raising new questions about how firms should price, deploy, and govern powerful technologies. Leading AI firms have positioned themselves as socially responsible entities committed to a dual mandate: generating profits for shareholders while enhancing social welfare and mitigating risks from AI. *OpenAI* adopted a capped-profit model. It allows investors to 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, including economic displacement, and give the labor market time to adjust.

How should firms with a social mandate price and deploy AI? Is a commitment to maximizing shareholders' 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 answers 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. In our formulas, optimal pricing is given by a *Modified Lerner's Rule*

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

where $\mathcal{M}$ summarizes the different 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 the cost of goods produced with these skills but disrupts the labor market of substitutable workers. The AI firm cares about profits, aggregate social welfare, and minimizing labor-market disruptions.

Our formula shows how the motives of socially-minded firms influence prices and quantities of AI in the short and long term. Profit motives push towards $\mathcal{M} = 1$, as in the traditional Lerner Rule. Aggregate social welfare considerations push towards $\mathcal{M} = 0$, or marginal-cost pricing. Setting

[1] See https://openai.com/index/openai-elon-musk/.

[2] https://www.anthropic.com/company

prices equal to marginal cost achieves the level of AI production and access that maximizes the size of the pie. These benefits are then weighed against distributional considerations, capturing who benefits the most from AI. Finally, the incentive to minimize labor market disruptions pushes for higher values of $\mathcal{M}$ that can exceed $\mathcal{M} = 1$ in the short run but not in the long run. This motive calls for a gradual and slow deployment path, where AI firm acts conservatively. This is because the cost of disrupting the labor market is higher in the short run, when workers have not had time to adjust, but decreases in the long run as the labor market adjusts. Our formula points to a tension between expanding access to AI (to maximize aggregate welfare) and the resulting loss in profits and labor market risks in the short run.

We conclude the paper with an empirical exploration, using US data. We use our formula to compute the optimal deployment path and prices of an AI capable of replacing human labor in each of 525 detailed jobs. That is, for every job in the US we imagine our tech firm develops an AI capable of replacing labor in that job at 50% the cost and ask how should a socially-minded firm price and deploy this AI. A *welfarist firm* that values both profit and welfare should price closer to marginal cost. For all jobs in the US, efficiency gains outweigh distributional concerns by a wide margin, since losses are not heavily concentrated at the bottom of the income distribution. 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. For every job, we also report optimal expansion plans for a firm that cares both about welfare and minimizing disruptions. So long as both incentives are equally important, we find that firms should price close to the profit-maximizing level in the short run and closer to marginal cost in the long run.

In several extensions we explore the role of the tax system and other forms of insurance between workers, AI that does not substitute for workers, and the role of competition between 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 prominent strand of this research 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) or by 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. 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 starts from Lerner's Rule and features an inefficient allocation in which too little AI is produced. In this context, expanding the use of AI yields first-order efficiency gains, which must be balanced against distributional and labor market stability objectives.

A third related literature studies the optimal deployment of AI accounting for existential risks (Jones, 2024, 2025) or uncertain risks that can be learned over time or via testing (Acemoglu and Lensman, 2024; Guerreiro, Rebelo and Teles, 2023). We view these papers as complementary and explore here a set of orthogonal issues related to welfare and labor-market disruptions due to AI.

Finally, our paper contributes to a large 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 as inputs into our numerical exercises.

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 radiology, copy-writing, journalism, customer service, and driving. These are all domains where AI systems can be trained to replicate human input. We also assume the technology is sufficiently advanced so that it can be operated autonomously and without input from workers. In our discussion section we extend our theory to account for the

possibility that AI is used for novel applications other than replicating human input.

1.1 The Economy

The economy is described in continuous time t. At each time, 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

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

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

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

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 has the capability to replicate labor input in a subset $\mathcal{A}$ of $\mathcal{S}$. The quantity of s input is

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

Here, q_{st} are units of AI-generated output, assumed 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 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

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

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

so that consumption of commodities (by people or computing facilities) 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 to say that financiers set the interest rate $r_t = \rho$ and determine the discount factor used by firms.

The objective of socially-oriented AI firms: the AI firm acts on 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 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. Note that investor welfare is already accounted for in profits, and so we do not include it once more to avoid double-counting it.

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

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

[3] In our formulation, the AI firm penalizes 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 extensions discuss a different formulation where firms minimize disruptions net of any indirect price benefits for workers.

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:

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

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

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. This motive in isolation calls for the usual monopolist pricing strategy of restricting quantities until its Lerner index $\mathcal{L} \equiv (P - MC)/P$ satisfies Lerner's Rule

$$\mathcal{L}_{st}\ =\ \frac{1}{\varepsilon_{st}}, \tag{2}$$

where $\varepsilon_{st} \geq 0$ is the (negative) of the demand elasticity of s at time t.[4]

The second term are *efficiency motives*. Because the firm cares about welfare, it has a motive to produce efficiently, increasing quantities until prices equal marginal cost $P = MC$ or $\mathcal{L}_{st} = 0$.

The third term are *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 are *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 $n_{st}^h > 0$ for at least some $s \in \mathcal{A}$.

The firm trades-off these motives optimally 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*

$$\mathcal{L}_{st} = \sum_{s'} \left((1-\mu) \frac{q_{s't}\ w_{s't}}{q_{st}\ w_{st}}\ +\ \mu \int_h g^h \frac{n_{s't}^h\ w_{s't}}{q_{st}\ w_{st}}\ dh\ +\ \lambda \int_{h: w_t^h < \bar{w}^h} \frac{1}{\bar{w}^h} \frac{n_{s't}^h\ w_{s't}}{q_{st}\ w_{st}}\ dh \right) \frac{1}{\varepsilon_{s'st}} \tag{3}$$

[4]This formula implicitly sets the cross demand elasticities to 0. More generally, the optimization results in the multi-product Lerner index as we discuss below.

where $\varepsilon_{ss't} \equiv -\frac{\partial \ln q_{st}}{\partial ln w_{s't}}$ *is the cross demand elasticity between* s *and* s'.

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 (3). □

To develop the intuition, we assume that the cross-demand elasticity is 0 for $s \neq s'$. The own demand elasticity ($s = s'$) is strictly positive and always remains in the formula. Then, the proposition shows that optimal pricing satisfies a Modified Lerner's Rule $\mathcal{L} = \frac{\mathcal{M}}{\varepsilon}$, where

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

The adjustment term deviates from 1 to balance firm objectives:

- The "1" is the usual profit maximization term.
- The "$-\mu$" pushes towards lower markups and higher quantities. This term reflects the firm desire to increase access to AI so as to raise 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 "$\lambda \int_h \frac{1}{\bar{w}^h} \frac{n_{st}^h}{q_{st}} dh$" is always positive and reduces quantities of AI produced. This captures the AI firm 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 an AI firm that only cares about profit ($\mu = \lambda = 0$), optimal prices are given by the usual Lerner Rule:

$$\mathcal{L}_{st} = \frac{1}{\varepsilon_{st}}.$$

For a *conservative* AI firm that also cares about minimizing labor market disruptions ($\mu = 0, \lambda > 0$), optimal prices satisfy

$$\mathcal{L}_{st} = \left(1 + \lambda \int_h \frac{1}{\bar{w}^h} \frac{n_{st}^h}{q_{st}} \, dh \right) \frac{1}{\varepsilon_{st}},$$

which exceed profit-maximizing prices. 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 are below profit-maximizing prices 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

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

For the general case with $\varepsilon_{ss't} \neq$ for $s \neq s'$, the firm's pricing also rule accounts for spillovers across skills through the production function.[5] If skills are domain-specific, then these are 0.

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 and (ii) prices and production and consumption plans that maximize households utility and firms profits (for firms producing commodities y). The characterization of the equilibrium in general is complicated, as the residual demand for AI q_{st} 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 $\mathcal{A}$.

[5] For example, when $\mu = \lambda = 0$, then we arrive at the standard, multi-product Lerner formula, which takes into account how increasing the quantity supplied of one good crowds demand in or out for other products.

(b) The utility function is $u_s(c_s) = \gamma_s^{1/\sigma_s}\ c_s^{1-1/\sigma_s}$, 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, as shown next.

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 (3)):*

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

and the demand curve (obtained from consumer demand):

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

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 that there is a unique equilibrium point where demand meets supply in all cases.

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

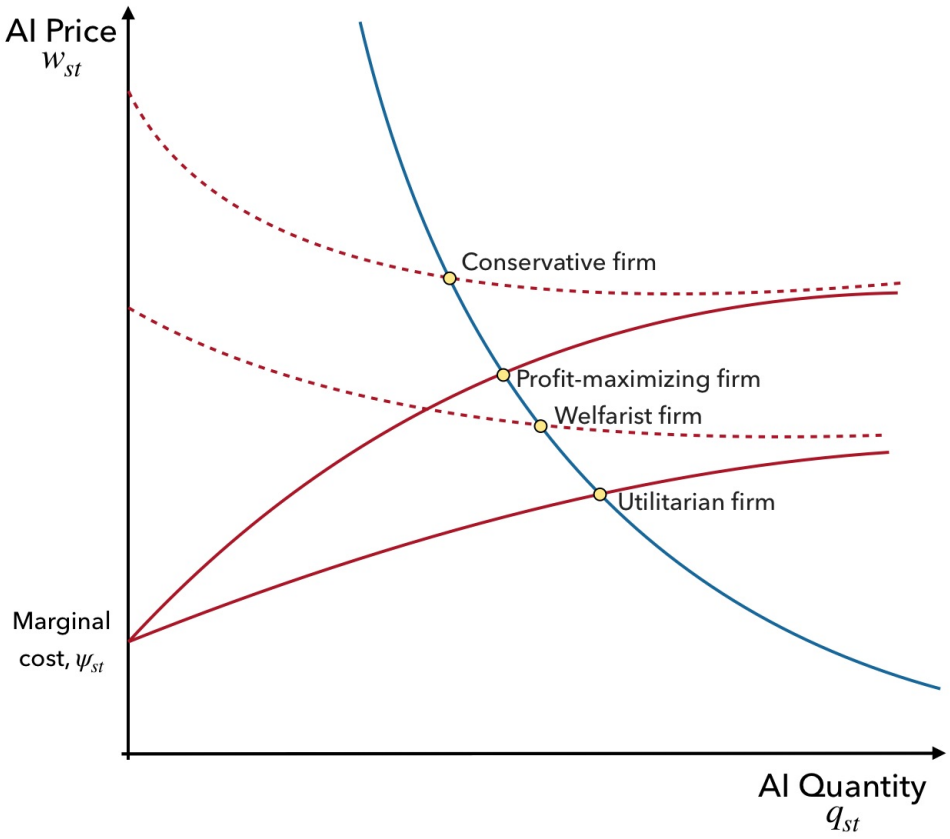

Figure 1: Equilibrium Supply and Demand for AI

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.

increase access and aggregate efficiency. The supply curve of welfarist and conservative firms are shifted upwards, reflecting incentives to restrict quantities so as to reduce harmful distributional or labor-market impacts of AI.

The figure also shows that the distributional motives of a welfarist firm or the stability motives of the conservative firm vanish as quantities rise. 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 shows that profit and efficiency motives scale with the quantity of AI used. Increasing the quantity of AI by 1% leads to a larger profit and efficiency increase if the use of AI is widespread. 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.[6]

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 and makes the residual demand faced by the AI firm more inelastic as time passes. This allows firms to set higher markups in the long run, pushing for a more frontloaded 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 in 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.

2 Scenarios for AI Transitions with Socially-Minded Firms

We now turn to a numerical exploration of the range of predictions generated by 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 cur-

[6]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 automation technology use becomes widespread.

rent cost. How should socially-minded firms deploy and price this technology? And what type of transition will we observe in practice if firms had different social objectives?

The formulas in the proposition show how to compute the optimal deployment path of any such AI for any skill s. In fact, answering this hypothetical question only requires taking the formulas in Proposition 2 to the data. The harder question of what are the specific skills for which this is more likely to happen does not need to be answered.

For our application, we focus on a firm that operates in the US economy and map skills to 525 detailed occupations appearing in 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).[7]

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 income distribution, assumed to have the same relative weight g^h.[8]

- 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 percentile h in occupation s.

[7] 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 occupations, Burstein, Morales and Vogel (2019) estimate an elasticity of substitution of 2.1 between 30 broad occupational groups. It makes sense to have a slightly larger value here since our occupational groups are finer.

[8] 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, given children a weight of 0.5 times 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.[9] 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. We see the 50% cost savings estimate as a reasonable bound on what the technology can do in some areas of the economy.

In the analysis, we contrast the optimal deployment plan of various firms. We consider a pure profit maximizer, a utilitarian firm ($\mu > 0$, $g_h = \lambda = 0$), a welfarist firm ($\mu > 0, g_h \neq 0, \lambda = 0$), a conservative firm ($\mu = g_h = 0, \lambda > 0$), and a multi-objective firm.

In the relevant scenarios, we set $\mu = 0.5$ and use the welfare weights g^h reported in Lockwood and Weinzierl (2016), who infer them 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 g^h are shown in Panel A of Figure 2. The values imply that the firm is willing to trade 1 dollar of profit for 1.9 dollars of value at the bottom of the income distribution and 3 dollars at the top.

Finally, in the relevant scenarios, we re-scaled λ by the average wage in the economy to ensure all terms have an equal scale and set $\lambda = 0.5$ in the relevant scenarios. This implies that the firm is willing to reduce profits by 1 dollar if this raises wages for the average worker by 2 dollars. These

[9] The factors behind this variable cost 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. Contrary to what some commentators believe, the variable computational and energy cost of using AI is not miniscule. If anything, these *inference costs* have been scaling as AI companies train larger and more complex models. Replicating human input can require multiple calls to these models, explaining why ψ_{st} can vary across jobs. One could also fold any residual labor costs (for example, of having a generic worker prompting the AI and verifying the output) into ψ_{st}.

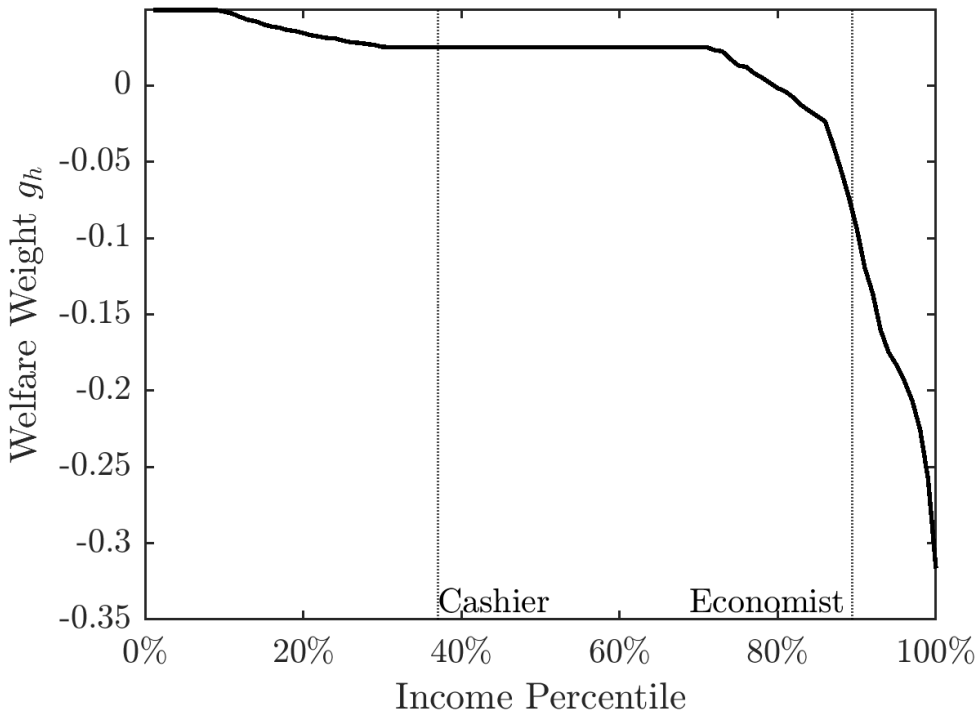

Figure 2: Welfare Weights Across the Income Distribution

Notes: The figure reports welfare weights g_h by income percentile. These are obtained as the welfare weights that rationalize the progressivity of the US tax system, and are from Lockwood and Weinzierl (2016).

are just scenarios that are meant to clarify how AI firms may act if they pursue a broader set of social objectives, but of course we do not know what is in their minds or hearts 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 skills.

As a benchmark, consider the case of the pure profit maximizer, in black. Markups for this firm at $t = 0$ range from 30% to 37% and are higher when its AI replaces more valuable skills. This is because high-pay skills are rare, limiting competition from human workers and increasing markups. As expected from our discussion of 2, markups rise at longer time horizons reflecting

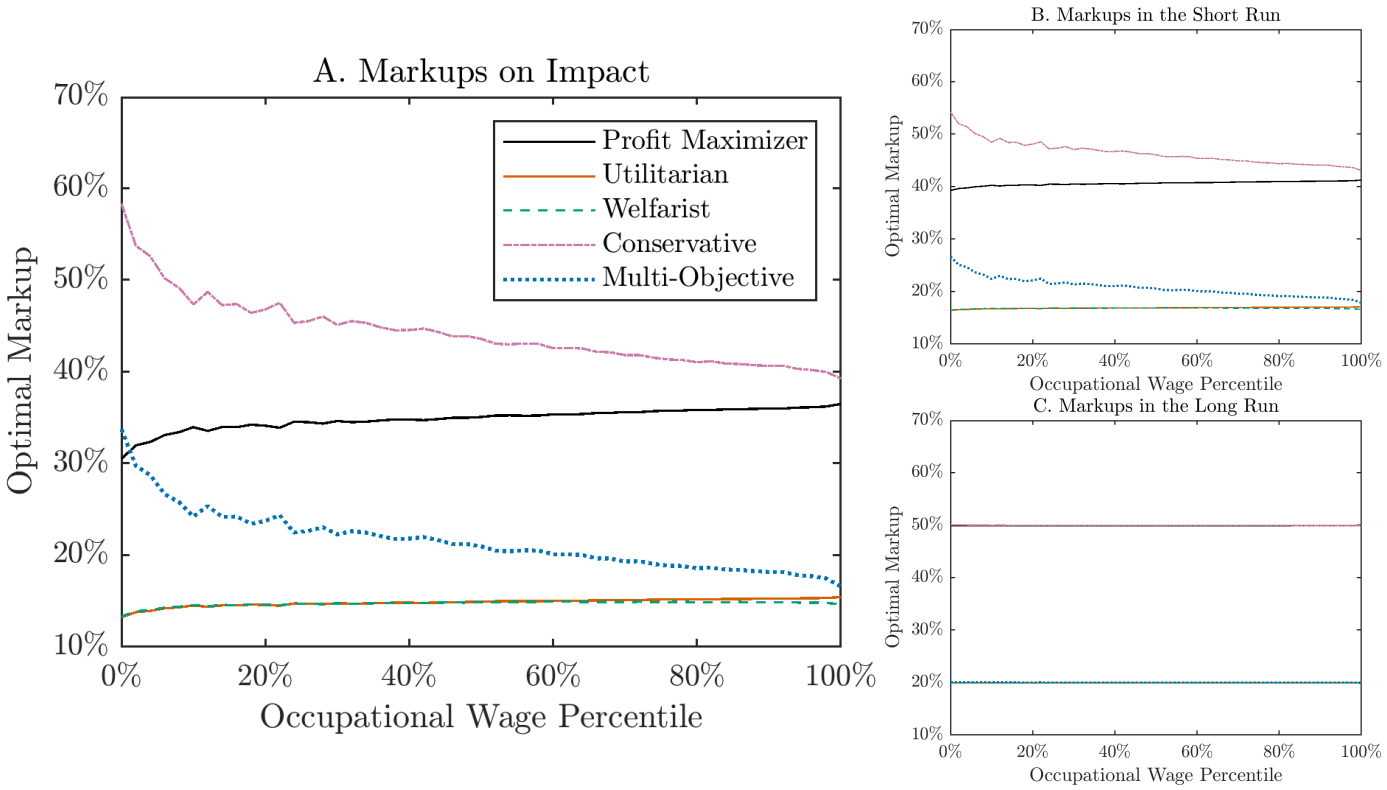

Figure 3: Equilibrium Markups in the Short and Long Run

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. Panel B and C report the same curves after 5 and 100 years.

reduced competition from workers as they reallocate over time. In the long run, the AI firm becomes the sole supplier of skill s and charges a common markup of $\sigma_s/(\sigma_s - 1)$ for any skill it can automate.

The utilitarian firm, in orange, charges lower markups than the profit-maximizing firm, ranging from 12% to 14% at $t = 0$ and converging to 20% in the long run. This is because the utilitarian firm has an incentive to lower prices below their profit-maximizing level. The lower prices expand access and increase aggregate efficiency.

The welfarist firm, in dashed blue, cares *both* about aggregate efficiency and distributional motives. The latter have a minor impact on equilibrium markups. They raise markups by less than .5% for AIs that automate low-pay jobs and lower markups for AIs that automate high-pay jobs by less than 1%. Moreover, in line with our discussion of Proposition 2, distributional motives vanish in the long run as people reallocate away from replaced skills. The fact that distributional motives play a muted role is one of the main findings of this section, discussed in detail below.

The conservative firm, in solid purple, balances profits against its concerns for labor-market stability. This firm ends up charging prices above the profit-maximizing level so as to minimize its labor market impact. This concern is particularly pronounced for AIs that automate low-pay occupations, since these generate more sizable labor-market disruptions. For this reason, equilibrium markups are higher (and quantities more restricted) for AIs that can automate low-paying occupations. However, in line with our discussion of Proposition 2, these sizable labor-market stability concerns vanish in the long run as people reallocate away from replaced skills.

Finally, the dashed blue line presents equilibrium markups for a multi-objective firm, which balances profit, efficiency, redistribution, and wage-stability concerns. This firm charges higher markups for AIs that can automate low-paying jobs, of roughly the same size as a pure profit maximizer. Instead, the multi-objective firm charges markups close to those of an utilitarian firm for AIs that can substitute high-pay occupations. In the long run, distributional and stability considerations vanish as workers reallocate. The multi-objective firm reduces markups at the bottom and charges a common markup of 20% so as to balance access and profits.

Figure 4 decomposes the role of each motive behind the optimal markup charged by the multi-objective firm. The dashed black line gives the contribution of profit motives, pushing for high markups, especially for high-pay jobs facing less competition from human workers. The orange line factors in aggregate efficiency considerations, which push for uniformly lower markups to balance profit against increased access. The green line factors distributional considerations, which play a small role as anticipated above. Finally, the blue line factors in wage stability concerns. These motivate the firm to restrict quantities and raise prices for AIs that can replace low-pay jobs, but have essentially no effect at the top.

Figure 5 complements the results by reporting equilibrium quantities. It shows the equilibrium quantity of the skill as a percent deviation from the baseline labor input in each occupation s. Quantities produced by AI range from one to four times that produced by workers at baseline and decline over time for firms without conservative motives as markups, and therefore prices, increase and vice versa. Lastly, the utilitarian firm generates the maximum increase in AI use.

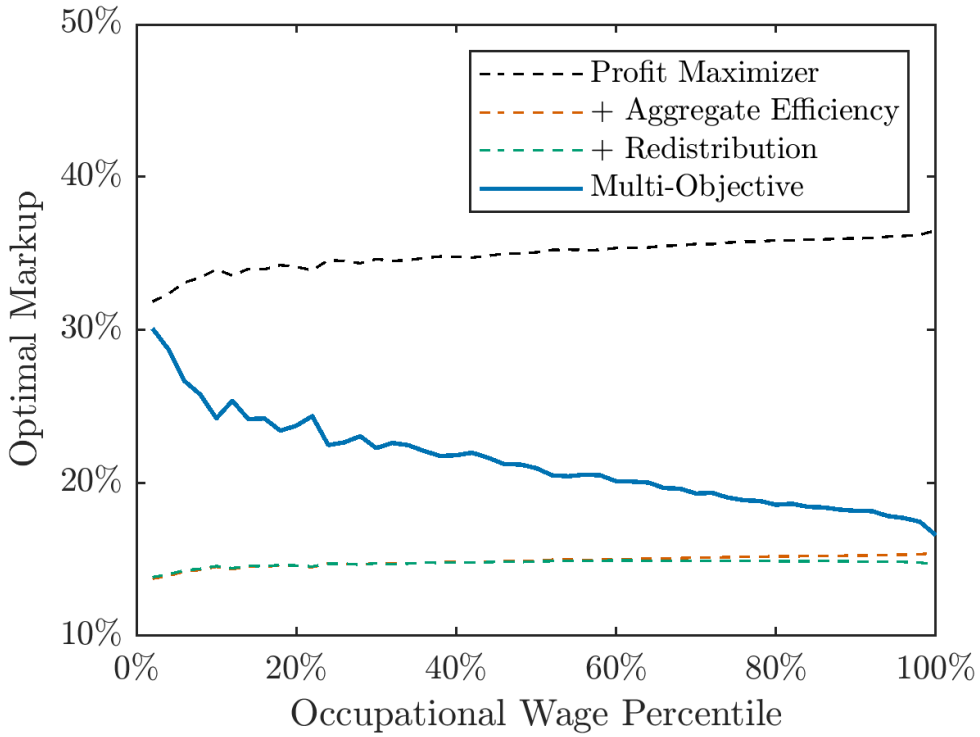

Figure 4: Decomposition of Motives Driving Markups Charged by Multi-Objective Firm

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 factors in 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. Panel B and C report the same curves after 5 and 100 years.

2.2 Why do distributional considerations play such a small role?

An important finding above is that a welfarist firm should prioritize the increase in access above distributional considerations. There are two reasons behind 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 4-times that supplied by workers at baseline. This pushes the firm to prioritize aggregate efficiency over distributional considerations.

More importantly, the relative welfare weights used do not generate a strong incentive to reduce existing inequalities by manipulating AI quantities. To see this, we compute the relative welfare weight associated with an increase in income of occupations as

$$g_s = \sum_h g_h \frac{\bar{n}_s^h}{\bar{n}_s}.$$

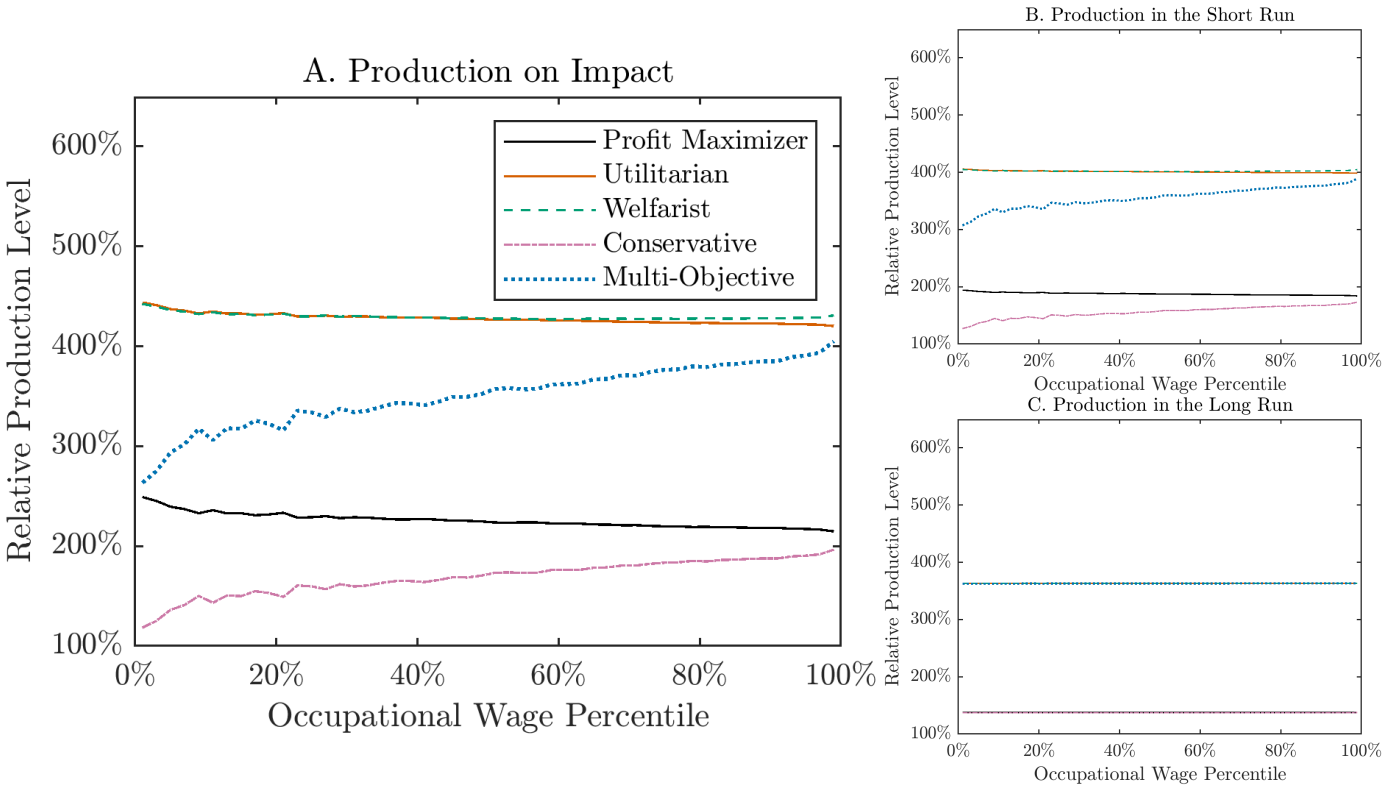

Figure 5: Equilibrium Quantities in the Short and Long Run

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. Panel B and C report the same curves after 5 and 100 years.

Figure 6 shows the relative welfare weight across the 525 detailed occupations in our data. Panel A reports the welfare weights and Panel B reports the average income percentile of people employed in each occupation. Welfare weights by occupation range from 0.05 for low-pay occupations to -0.15 for high-pay ones. For example, *Cashiers*—a job employing households that on average come from the 37 percentile of the income distribution—has a small relative weight of 0.05. *Economists*—a job employing households that on average come from the 82 percentile of the income distribution—has a negative relative weight of -0.15. These differences between occupations at polar ends of the pay scale are small and do not justify meaningful deviations in the way an AI capable of replacing economist relative to one capable of automating cashiers.

To illustrate this point, consider an AI firm that places ten times more weight on distributional considerations than the weight inferred from the US tax system. For this AI firm, welfare weights g^h are ten times those in Figure 2 and skill-specific weights (proxied by occupation) are also ten

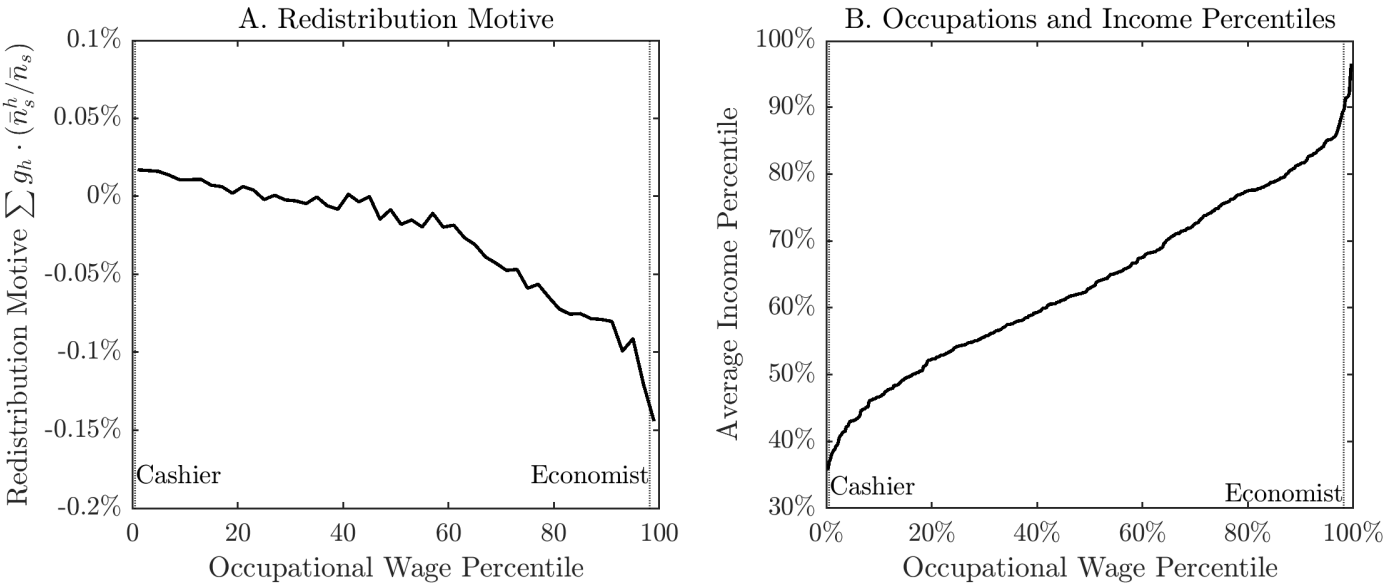

Figure 6: Welfare Weights Across Occupations

Notes: Panel A plots welfare weights g_s by skill (proxied by occupations). These are obtained as $g_s = \sum_h g_h \frac{\bar{n}_s^h}{\bar{n}_s}$, where the individual weights g^h are from Lockwood and Weinzierl (2016). Panel B plots the average income percentile of workers within an occupation against the percentile of the average wage in the occupation among all occupations.

times those in Figure 6. Equilibrium markups for this firm, relative to a utilitarian one, are shown in Figure 7. The left panel shows that a firm with such strong distributional motives should charge 6% lower markups for AI systems capable of replacing high-wage jobs and 2% higher markups for AI systems capable of automating low-pay jobs, relative to the utilitarian firm. The right panel shows that this policy results in a reduction of 4% in the equilibrium quantity of AI automating low-pay jobs and an 12% increase in the quantity of AI replacing high-pay jobs.

In sum, our results suggest that distributional considerations, even ten times as pronounced as those implicit in the US tax system, do not appreciable change the way a firm should deploy its AI technology. A firm that seeks to maximize welfare should focus instead on balancing profit against aggregate efficiency by lowering prices and increasing access. From a social point of view, this is already close to being the best they can do. If anything, distributional concerns push for broadening access the most for AIs that can automate high-pay jobs.

2.3 Should more productive AI be priced differently?

Our baseline results considered the optimal deployment of AIs capable of replacing human skill at 50% the base cost. Should extremely productive AIs be priced differently? Suppose for example

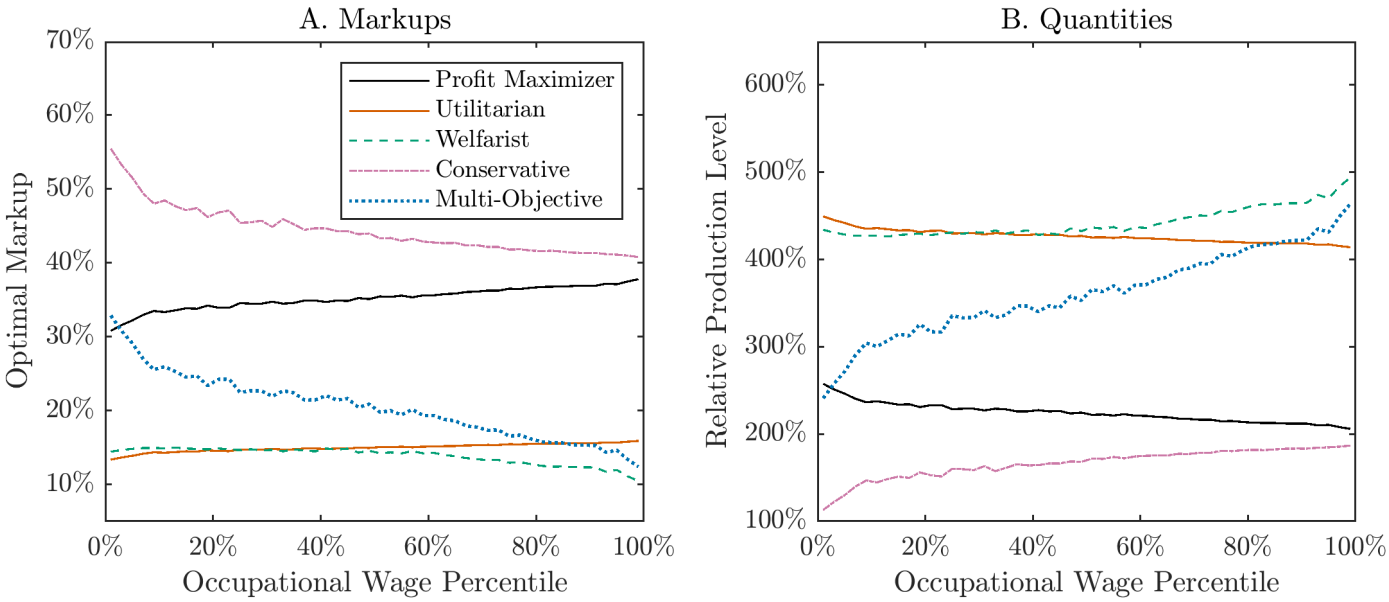

Figure 7: Markups and Quantities on Impact for Stronger Redistributive Preferences

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 report optimal markups and Panel B the associated quantities for the automated skill.

that $\psi_s = .2\ \bar{w}_s$, so that AI can replicate human labor at 20% the cost in automated jobs.

Figure 8 reports equilibrium prices and quantities for such AIs across occupations at $t = 0$. Relative to our baseline, markups are slightly larger and quantities and 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. This effect is especially pronounced in lower-wage occupations, leading to more uniform markups.

More importantly, the figure shows that both distributional and labor-market stability motives are weaker for more productive AIs. This can be seen from the fact that the equilibrium outcomes for conservative firms are now close to those of a pure profit maximizer, while 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 motives do not. The lesson is that as firms develop more productive and less costly AIs, profit maximization or efficiency considerations become dominant. A socially-minded firm with a sufficiently productive AI should behave essentially as an utilitarian one and prioritize a balance between maximizing access and increasing profit. Distributional and stability considerations should take a back seat at that point.

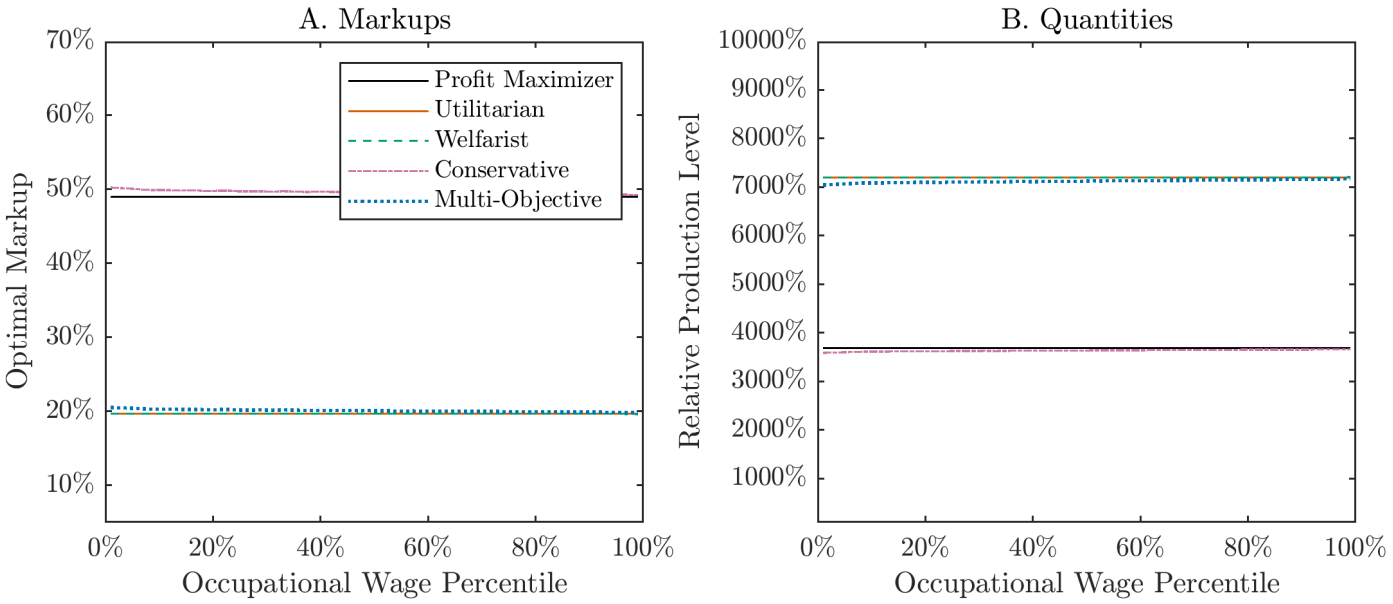

Figure 8: Markups and Quantities on Impact for More Productive AI

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

2.4 Do these conclusions apply to occupations with the highest risks?

As discussed above, our approach ignores the harder question of what are the specific skills and occupations that will be substituted by AI in future years. Instead, we focus on exploring how optimal pricing should vary across all potential jobs.

We now explore whether the incentives for redistribution and stability are stronger for occupations with the highest risk of automation by AI. 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 9 shows that highly exposed occupations are quite *average* and do not carry different distributional and stability considerations from others. The left panel shows that highly exposed occupations have average welfare weights g_s on the vicinity of zero. The right panel shows that highly exposed occupations carry an average stability consideration, reflecting the fact that there are occupations at risk over the entire income distribution. 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.

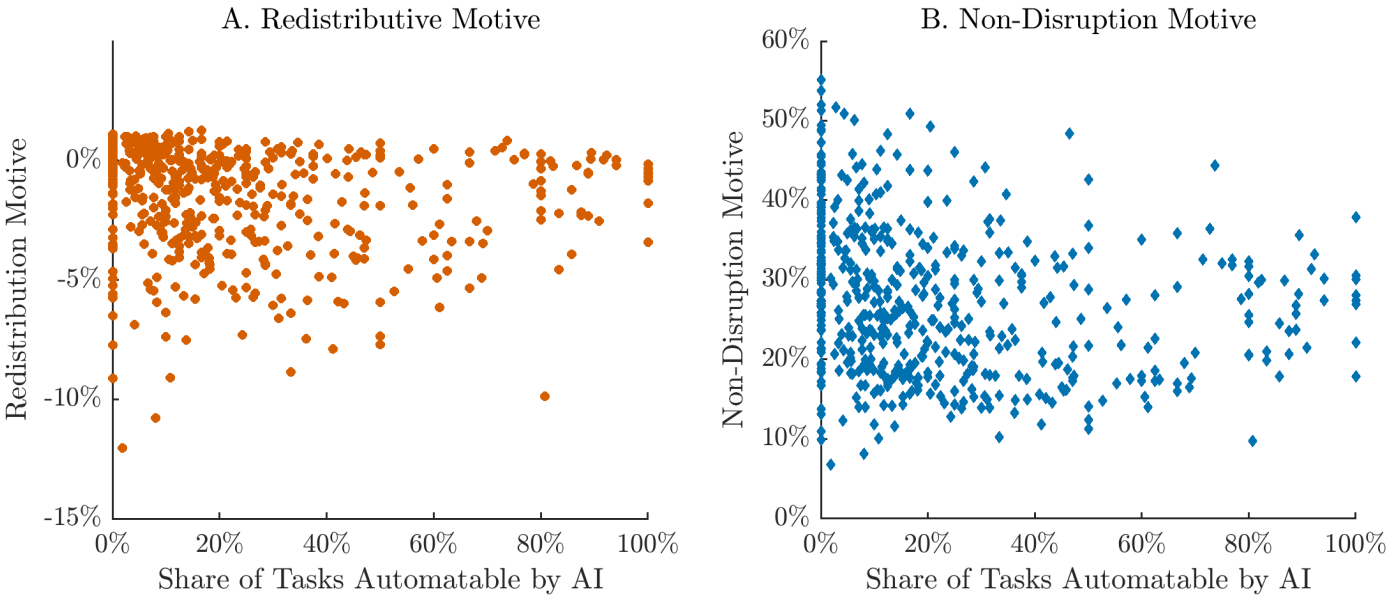

Figure 9: Welfare Weights and AI Exposure

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

3 Theoretical 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 there are multiple AI firms competing a la Cournot. All else equal, adding competition lowers optimal markups based on profit maximizing and utilitarian motives as firms' elasticity of demand increases. However, the strength of any welfarist or conservative motives for redistribution are perserved.

3.1 Taxes

Our formulas above ignored the mitigating force of the tax system, which provides some level of insurance and redistribution from winners to losers. To account for this, assume the after-tax labor income of household h is

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

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 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*

$$\mathcal{L}_{st} = \left(1 - \mu \; + \; \mu \int_h g^h \; \mathcal{T}'(w_t^h) \; \frac{n_{st}^h}{q_{st}} \; dh \; + \; \lambda \int_h \frac{\mathcal{T}'(w_t^h)}{\mathcal{T}(\bar{w}^h)} \; \frac{n_{st}^h}{q_{st}} \; dh \right) \frac{1}{\varepsilon_{st}} \tag{6}$$

It follows that higher levels of redistribution via lower $\mathcal{T}'(w_t^h)$ reduce weight on both distributional and stability concerns. In the extreme case of full redistribution, i.e., $\mathcal{T}'(w_t^h) = 0$, both distributional and stability considerations vanish, allowing the AI firm to focus exclusively on balancing profits vs aggregate efficiency by expanding access.

3.2 AI as creating new goods and services

Our formulas adopted the natural position that AI technologies can replicate human skill in some areas of the economy. This is to be expected, since AI systems are trained in human-generated data to learn how to mimic what we already know how to do. However, some argue that large systems trained in vast amounts of data might develop new capabilities and figure out ways to produce goods and services that have so far escaped all of us.

A simple way of thinking of this type of AI is by assuming it introduces a new skill into the economy $s \notin S$. In the example economy in Section 1.2, this means that household utility is now

$$u^h(c) = c_{0t}^h + \sum_{s \in \mathcal{S}} \gamma_s^{1/\sigma_s} \; c_s^{1-1/\sigma_s} + \gamma_s \; \; c_s^{1-1/\sigma_s},$$

for $\sigma_s > 1$. This is a simple way of capturing what it means to create an AI that performs tasks or

has capabilities that are unlike anything humans do.

Proposition 4. *In an economy where (a)–(c) hold, the optimal pricing of an AI with novel skills not in $\mathcal{S}$ satisfies a Modified Lerner's Rule*

$$\mathcal{L}_{s't} = \left(1 - \mu\right) \frac{1}{\sigma_{s'}}. \tag{7}$$

For this class of AIs, socially-responsible firms should price closer to MC and increase access. There are no distributional trade-offs nor labor-market disruptions to worry about.

3.3 Competition among AI producers

Finally, given the recent advances by multiple AI companies, it is natural to consider competing firms. To simplify the exposition, we assume that there are M_s identical AI firms capable of producing q_{st} that compete in quantities a 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*

$$\mathcal{L}_{st} = \left(\frac{1 - \mu}{M_{st}} + \mu \int_h g^h \frac{\bar{n}_s^h}{q_{st}} e^{-\alpha t} \, dh + \lambda \int_h \frac{1}{\bar{w}^h} \frac{\bar{n}_s^h}{q_{st}} e^{-\alpha t} \, dh \right) \frac{q_{st}}{q_{st} + \bar{n}_s \, e^{-\alpha t}} \frac{1}{\sigma_s} \tag{8}$$

The formula shows that the equilibrium mimics what we would get with a single firm that now places more weight on aggregate efficiency considerations. This makes intuitive sense, since competition forces firms to price closer to their marginal cost.

One interesting aspect is that distributional and stability considerations are not diminished by competition. As an example, assume distributional motives are positive. A large number of socially minded firms would then reach an equilibrium where prices remain bounded away from marginal costs, reflecting their distributional and stability considerations. The reason is that there are two offsetting effects. On the one hand, firms realize their demand is more elastic, pushing them to charge lower markups. On the other hand, more competition means each firm gets smaller. In the limit, firms are so small that the efficiency or profit costs of raising markups becomes tiny and they forego these objectives to prioritize their other motives.

4 Conclusion

How should firms act when entrusted with transformative technologies like AI and a social mandate? This paper provides a formal framework to address that question by extending Lerner's Rule to account for the broader objectives of socially-minded firms: 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 U.S. data across hundreds of occupations, we find that utilitarian firms—those valuing both profits and efficiency—should price close to marginal cost. The resulting gains in aggregate welfare generally outweigh distributional concerns. In contrast, conservative firms that prioritize labor-market stability should set higher markups in the short run, particularly in sectors where AI may displace low-wage workers. As workers reallocate and labor-market disruptions subside, firms can gradually reduce prices and expand access.

These results reveal a fundamental trade-off facing socially-minded firms: expanding access to AI promotes efficiency but intensifies short-run labor-market risks and reduces profits. Our framework quantifies this trade-off and provides transparent pricing principles to help guide both firm behavior and policy design in a period of rapid technological change.

Ultimately, these insights contribute to a broader conversation about how socially responsible deployment should proceed and how firms' own social abitions and ideals shape the roll-out of technologies with potentially large societal impacts.

References

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.

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.

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.

Toner-Rodgers, Aidan. 2024. "Artificial Intelligence, Scientific Discovery, and Product Innovation."

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

A Proofs

To derive (1) observe that following an arbitrary change in quantities by the AI firm, we get

$$\delta S = \int_0^t e^{-\rho t} \Big\{ (1+\epsilon\mu) \sum_s \left(q_{st}\ \delta w_{st} + \left(w_{st} - \psi_{st} \right)\ \delta q_{st} \right) + \int_h \mu^h \left(\sum_s n_{st}^h\ \delta w_{st} - \sum_j c_{jt}^h\ \delta p_{jt} \right) dh + \lambda \int_h \frac{1}{\bar{w}^h} \sum_s n_{st}^h\ \delta w_{st}\ dh \Big\}\ dt.$$

The first line gives the changes in profits flowing to financiers, which receive a weight $1 + \epsilon\mu$. The second line gives the change in households utility. By assumption, $c_{0t}^h > 0$. This implies that the marginal value of income for all households in period t is $e^{-\rho t}$. The change in household utility is then given by the second line, where the change in household net income arising from price changes is added with a weight $e^{-\rho t}$, capturing its marginal value, times μ^h, capturing its social value. 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 arising due to price changes shows up. The third line gives the effects via labor market disruptions, which are assumed to be a function of wages.

This can be rewritten as

$$\delta S = \int_0^t e^{-\rho t} \Big\{ (1+\epsilon\mu-\bar{\mu}) \sum_s \left(q_{st}\ \delta w_{st} + \left(w_{st} - \psi_{st} \right)\ \delta q_{st} \right) + \bar{\mu} \left(\sum_s \left(q_{st}\ \delta w_{st} + \left(w_{st} - \psi_{st} \right)\ \delta q_{st} \right) + \int_h \left(\sum_s n_{st}^h\ \delta w_{st} - \sum_j c_{jt}^h\ \delta p_{jt} \right) d \right. + \bar{\mu} \int_h g^h \left(\sum_s n_{st}^h\ \delta w_{st} - \sum_j c_j^h\ \delta p_j \right) dh + \lambda \int_h \frac{1}{\bar{w}^h} \sum_s n_{st}^h\ \delta w_{st}\ dh \Big\}\ dt.$$

or

$$\delta S = \int_0^t e^{-\rho t} \left\{ (1 + \epsilon\mu - \bar{\mu}) \sum_s \left(q_{st}\ \delta w_{st} + \left(w_{st} - \psi_{st} \right)\ \delta q_{st} \right) \right.$$
$$+\ \bar{\mu} \left(\sum_s q_{st}\ \delta w_{st} + \sum_s \left(w_{st} - \psi_{st} \right)\ \delta q_{st}\ +\ \sum_s n_{st}\ \delta w_{st} - \sum_j y_{jt}\ \delta p_{jt} \right)$$
$$+\ \bar{\mu} \int_h g^h \left(\sum_s n_{st}^h\ \delta w_{st} - \sum_j c_{jt}^h\ \delta p_{jt} \right) dh$$
$$\left. +\ \lambda \int_h \frac{1}{\bar{w}^h}\ \sum_s n_{st}^h\ \delta w_{st}\ dh \right\} dt.$$

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

$$\sum_j y_{jt}\ \delta p_{jt} - \sum_s (q_{st} + n_{st})\ \delta w_{st} = 0.$$

This implies

$$\delta S = \int_0^t e^{-\rho t} \left\{ (1 + \epsilon\mu - \bar{\mu}) \sum_s \left(q_{st}\ \delta w_{st} + \left(w_{st} - \psi_{st} \right)\ \delta q_{st} \right) \right.$$
$$+\ \bar{\mu} \sum_s \left(w_{st} - \psi_{st} \right)\ \delta q_{st}$$
$$+\ \bar{\mu} \int_h g^h \left(\sum_s n_{st}^h\ \delta w_{st} - \sum_j c_{jt}^h\ \delta p_{jt} \right) dh$$
$$\left. +\ \lambda \int_h \frac{1}{\bar{w}^h}\ \sum_s n_{st}^h\ \delta w_{st}\ dh \right\} dt.$$

To conclude, note that $c_{jt}^h = y_{jt}$ because of quasi-linearity. The term $\sum_j c_{jt}^h\ \delta p_{jt}$ is then common to all households and cancels because $\int_h g^h\ dh = 0$. This simplification yields (1).