What Should We Tax When We Tax AI?*

Giacomo Brusco
Francesco Capozza

This version: October 12, 2026

Abstract

Generative AI is a metered intermediate input: concentrated in supply, powered by imperfectly priced electricity, and substitutable for some labor but not all. Four tax proposals track these features: a token tax, an excess profits tax, an electricity levy, and a robot tax. We embed all four in one task model with Roy sorting and endogenous entry. The optimal ad valorem tax on usage has three terms, all zero at the Diamond–Mirrlees benchmark. Predistribution is scaled by an exposure gap the literature sets to one; across seven public measures no value exceeds 0.45, and all four direct AI measures give a *negative* gap. The Pigouvian term is a few percent for a token against sixty percent or more for a kilowatt-hour. The market power term enters negatively and dominates: under the constant-elasticity demand the model generates, the efficient usage wedge for a domestically owned sector is negative at any markup above about ten percent, whether or not entry is free. That sign turns on pass-through: if providers absorb part of a tax, it extracts rent and the optimum turns positive.

Keywords: Generative AI; intermediate inputs; production efficiency; Pigouvian taxation; rent taxation; automation.
JEL: H21, H23, Q58, O33.

*Brusco: UB, IEB: gbrusco@ub.edu; Capozza: UB, IEB, CESifo, IZA@LISER: francesco.capozza@ub.edu. We do not report any conflict of interest.

1 Introduction

Generative AI enters the economy as an input to cognitive work. Earlier automation waves substituted for codifiable tasks, which made the incidence question relatively clean: middle-skill routine occupations lost ground, and the job-polarization literature traced the consequences. Modern systems learn by example, acquire knowledge that resists codification, and scale across compute instead of across people (Ide and Talamàs, 2025), so they reach work no earlier technology reached. While the productivity gains are real and measurable; their incidence is not settled, and Section 2 shows diffusion is narrower and its distributional direction more ambiguous than the tax debate assumes. Technological innovation with consequences so uncertain and potentially pervasive is a poor candidate for an instrument that assumes it knows who wins and loses.

Four proposals to tax artificial intelligence are on the table. A levy on usage, charged per token or per call (Korinek and Lockwood, 2025). An excess profits tax on the firms that sell it, or its voluntary cousin, a commitment to share windfall profits (O'Keefe et al., 2020). A charge on the electricity its data centers consume, or a tariff that makes large loads pay the grid costs they cause (Jacobs, 2025; Riu et al., 2025). And a robot tax, levied on AI because it displaces workers (Guerreiro et al., 2022; Thümmel, 2023). They are usually discussed as variations on a single impulse, that a fast-growing and profitable sector should pay something, but they are aimed at different problems and the problems do not point the same way. A tax on tokens does nothing about the carbon intensity of the grid serving a data center; a tax on profits does nothing about displacement. And one of the four runs backwards: if AI is sold above marginal cost it is consumed too little, and correcting that calls for a lower price. Whether a tax on AI usage should be positive at all is therefore not obvious, and cannot be settled by adding up arguments developed in isolation.

The natural benchmark is that AI inputs should be untaxed. Diamond and Mirrlees (1971) show that when the government can tax final goods and incomes optimally, production efficiency should be preserved and intermediate inputs left untaxed, whatever their distributional consequences downstream. The benchmark is not wrong. It is conditional, and its conditions are the four features that make this case interesting: a missing market for the environmental cost of the electricity, market power in supply, a redistributive motive the income tax cannot fully serve, and an open economy in which some of the rent accrues abroad. Each of the four is a reason the benchmark might fail, and none has been priced against the others in a single model. This paper builds that model and prices them. AI is a metered service produced from compute and electricity by a sector

with a markup, used in tasks that would otherwise be performed by labor, with workers sorting across tasks on comparative advantage and free but costly entry into provision. The government has a linear income tax, a usage tax, a cash flow tax, and a Pigouvian charge at the meter, but it cannot tag workers by type.

The optimal ad valorem tax on AI usage[1] is the sum of three terms, one for each coherent argument in the debate:

$$t_U^\star \;=\; \underbrace{\text{predistribution}}_{\text{wage compression}} \;+\; \underbrace{\text{Pigouvian}}_{\text{external cost / price}} \;-\; \underbrace{\text{market power}}_{\text{underprovision}}.$$

All three vanish at the Diamond–Mirrlees benchmark, and each is a sufficient statistic that can in principle be measured. Three findings follow.

First, the externality is real but it is on the wrong base. The uninternalized external cost of a unit of AI is a few percent of the price of a token, dominated by unpriced carbon; the same cost per kilowatt-hour at the data center is sixty percent or more of the price of that electricity. The harm-to-value ratio varies tenfold across workloads, so no single ad valorem rate on usage can track it – not to mention model training. A per-kilowatt-hour charge tracks electricity consumption by construction and reaches an efficiency margin a usage tax cannot touch.

Second, the redistributive case is weaker than the robot-tax literature implies, for a reason that literature cannot express. It identifies the group a planner wants to protect with the occupation the technology reaches, which fixes at one what we call the *exposure gap*: the difference in the share of tasks exposed to AI substitution for two different groups. We model task assignment as a Roy choice, which avoids assuming workers are stuck conducting the same tasks. We measure the exposure gap using several measures available in the literature and find nothing above 0.45 – less than half than the value implied by unadjustable task assignment. Indeed, all four direct AI measures of the exposure gap yield values below zero, further weakening the redistributive case. This is for an intuitive reason: while robot exposure is concentrated in codifiable, lower-skill work, generative-AI exposure is concentrated in cognitive work done by the highly educated. Dispersion of comparative advantage is also not a mobility parameter: it governs sorting and switching at once and in opposite directions, so a world in which the losing group is sharply identifiable is one in which nobody is near the switching margin.

Third, the market-power term is negative and, under the demand system the model

[1]We do not consider excise taxes on A.I. as they are well known equivalent, in case of perfect competition, or worse, in case of imperfect competition, than ad valorem commodity taxes (Keen, 1998).

generates, decisive. With the number of providers fixed, training costs are sunk and any margin over marginal cost results in underprovision. The market power term is further weighted by the planner's weight on shareholders, which we derive by requiring consistency with the weight on routine workers.[2] Under free entry profit is zero, so ownership is irrelevant, and the term instead reflects how a tax raises the markup by driving out entrants. In this case, the gain from cheaper AI then accrues to buyers. In both regimes the term exceeds the other two at any markup above about ten percent, and the markup the model can support is itself bounded: no firm faces less elastic demand than a collusive cartel would.

Netting the three at our central calibration, the efficient usage wedge for a domestically owned sector is about -15 percent with entry blocked and -10 percent under free entry, and it is negative in 97 percent of draws in either regime; for an importer with entry blocked it is about +6 percent. *Rates under debate sit at or above the rates that are optimal under the model's demand system.* The sign of the optimal tax rate is the least robust result, and we can characterize what determines it: it turns on pass-through. When providers absorb part of a tax, the tax extracts rent, and with linear demand calibrated to the same markup the domestic optimum becomes positive, about eight percent.[3] Setting the market-power term aside entirely, the remaining terms give a wedge of about 6 percent.

One further element should be considered in thinking about the optimal tax system around AI. Compute is mobile, so a unilateral charge on data center electricity relocates load rather than abating it, and a destination-based levy on AI *consumption* is the backstop. That argument is about leakage, not robots or rents, and rests on a parameter nobody has estimated yet.

We intend the expression diagnostically rather than as a menu of legislation: it says which instruments are dominated and thinks through which measurements would change the answer. Besides the degree of leakage, three matter. First, the markup over marginal cost and, second, the pass-through of taxes and costs into AI prices decide the sign of the optimal usage tax; both can be measured from price responses, and neither has been reported for AI services. The exposure gap fixes the redistributive term, and has never

[2]Requiring consistency with how one weights routine workers limits how much shareholders can be discounted: a planner only mildly averse to inequality discounts shareholders only mildly.

[3]If no one switches tasks, which is the case the robot-tax literature assumes, the predistribution term rises to its fixed-assignment value of 7.3 percent; with entry blocked the domestic wedge is then -11.7 percent and an importer's $+10.3$ percent. The fixed-assignment value requires perfect sorting, at $\Delta_\ell = 1$, $\mathcal{E} = 0$; in the Roy model reallocation also vanishes when dispersion is wide, but there the exposure gap vanishes too.

been reported because the literature sets it to one by assumption.[4]

The paper sits at the intersection of four literatures. From *intermediate-input taxation* we take the Diamond–Mirrlees benchmark and the two classical routes around it, Naito (1999) and Sandmo (1975), whose additivity of Ramsey and Pigouvian terms yields the structure of our main result.[5]

From the *optimal taxation of automation* we take the robot-tax formulation of Costinot and Werning (2023), Thümmel (2023) and Guerreiro et al. (2022), whose routine and non-routine occupational choice is closest to our own, and depart from all three on the exposure gap as described above. Kleven and Zidar (2026) show why such machine taxes arise at all. In a task-based model in which labor responds on the extensive margin and each occupation pays a distinct wage, the income tax is effectively occupation-specific, the partial-equilibrium tax formulas hold exactly in general equilibrium, and the case for taxing machines disappears. Our predistribution term is the channel they close: it exists because the income tax cannot condition on the task a worker held, and the exposure gap measures how much of it survives when it cannot. Rothschild and Scheuer (2013) supply the template for optimal taxation with Roy sorting.

The distributional question has two close precedents. Skill-biased and routine-biased technical change raised the return to education by substituting for codifiable work (Autor et al., 2003), and its implications for taxation are worked out by Ales et al. (2015) and Löbbing (2024). Some have argued the China shock concentrated losses on identifiable local labor markets (Autor et al., 2013; Galle et al., 2023), and Costinot and Werning (2023) treat trade and automation as one sufficient-statistic problem. Both literatures could credibly take the direction of the shock as given, since the technologies and trade flows they studied fell most heavily on less-educated workers. Our exposure gap measures the same object for AI and finds that its sign cannot be taken as given (Section 6). That changes how losers should be compensated: when the sign is uncertain, transfers should condition on realized outcomes rather than on predicted exposure, since a technology tax built on the wrong sign redistributes toward the better-off (Section 8).

From the *economics of AI and energy* we take the positive facts—International Energy Agency (2026), Shehabi et al. (2024), Luccioni et al. (2024) on load and energy per query, and the contested evidence on grid cost allocation—and supply the normative half, which is largely missing. From the *taxation of the digital economy* we link the design

[4]None of the three requires new econometrics. We compute the exposure gap in Section 6.

[5]Scheuer and Werning (2016) show that in exactly the assignment and Roy models we use, general-equilibrium wage effects wash out of the optimal *nonlinear* income tax, so our redistributive term should be read throughout as an upper bound generated by a restricted instrument set.

of destination-based and cash flow instruments (Auerbach and Devereux, 2018) to the current debate: a cash flow tax on rents and a turnover tax on revenue are substitutes rather than complements, one neutral and one distortionary; our argument cuts in favor of the first even though much of the public debate has focused on the latter. To the *industrial organization of taxation* we apply the pass-through amplification that establishes for free-entry oligopoly, characterizing it for this setting in Proposition 6: with free entry and fixed costs an ad valorem tax on a concentrated sector raises the equilibrium markup by driving out entrants, so its burden on buyers exceeds the revenue it raises.[6]

Section 2 sets out the four claimed departures and the evidence for each, Section 3 derives the optimal usage tax, Section 5 reads the three terms, Section 6 calibrates and measures the exposure gap, and Section 7 scores the four proposals. Proofs and extensions are in the appendices.

2 Four claimed departures from the benchmark

The case for treating AI differently from other intermediate inputs rests on four claims. Each corresponds to a condition of the production efficiency theorem, and each is a claim about the world rather than about the model, so we state the evidence for each before building anything on it.

An ambiguous distributional direction. The same model can compress or widen the earnings distribution depending on whether it advises a worker or replaces one. Autor (2024) makes the case for compression: AI supplies expert judgment instead of executing rules, so it can extend high-stakes decision work to people who lack the credential for it today, and Autor and Thompson (2025) show the answer depends on whether the technology takes over the low-expertise or the high-expertise part of a job. Diffusion is narrow as well as ambiguous. Automating an automatable task may not be worth the compute it consumes, in which case adoption stops well short of exposure. The Census Bureau's firm survey finds AI in use at 18 percent of firms, 32 percent weighted by employment, with most adopters using it in three or fewer business functions (Bonney et al., 2026). Two thirds of firms reporting any task effect report augmentation alone,

[6]The comparative statics of commodity taxation under imperfect competition are well developed: Delipalla and Keen (1992) compare ad valorem and specific taxation in oligopoly with and without free entry, and Auerbach and Hines Jr. (2001) characterize optimal corrective policy when the degree of competition is uncertain, obtaining the same trade-off between subsidizing output in concentrated sectors and raising revenue elsewhere that drives our market-power term.

AI-driven employment declines appear at 2 percent of firms, and 16 percent of adopters report replacing existing software or equipment with AI. Net employment counts can miss labor substitution offset by reassignment, so these figures do not settle whether capital or labor is displaced more often, but they do not describe a technology whose losers are easy to identify. Section 6 measures occupational exposure and finds that its direction depends on the measure: routine-displacement measures point the way the robot tax assumes, while every direct AI measure points the other way.

A missing redistributive instrument. Governments cannot observe whether a given worker is a substitute or a complement for AI, so they cannot levy a tax that depends on it. When exposure is unobservable, manipulating relative factor prices can provide an additional redistributive margin by changing equilibrium wages. This is related to the Naito (1999) channel, in which commodity taxes relax the incentive-compatibility constraints facing an optimal nonlinear income tax. Our mechanism is more restrictive: because we limit the government to a linear income tax, the wage channel remains valuable, whereas under the unrestricted nonlinear schedule studied by Scheuer and Werning (2016) it disappears. This is the robot-tax argument, and it is the oldest and least contested of the four departures from the benchmark. Section 5.3 shows the term collapses to zero under a fully nonlinear schedule. AI is a particularly relevant case for this mechanism because, unlike many conventional inputs that primarily complement workers, it can both complement labor and substitute for it in specific cognitive tasks. The balance between these forces is an empirical question, but the substitution margin is what gives the redistributive claim its potential relevance and motivates the task structure of Section 3.

A missing market. The electricity data centers consume carries two costs its price does not fully reflect. The first is environmental. At a social cost of carbon of $190 per ton (U.S. Environmental Protection Agency, 2023) and a grid intensity of 0.4 kg per kWh serving U.S. data centers (International Energy Agency, 2025), unpriced carbon damage alone is roughly $0.076 per kWh where no carbon price applies—close to the entire industrial electricity price of about $0.08.

The second is grid cost shifting. Under cost-of-service regulation a utility recovers costs across customer classes on the principle of cost causation: each class should pay for the infrastructure built to serve it (Borenstein and Bushnell, 2022). Large new loads that require transmission upgrades and generation capacity strain that principle, because the incre-

mental cost is lumpy and the allocation rules were not written with hundred-megawatt single customers in mind. Whether shifting has actually occurred is genuinely disputed—Energy and Environmental Economics (E3) (2026) finds large-load tariffs largely recover incremental cost, while Jacobs (2025) and Monitoring Analytics (2025) find substantial socialization—and Section 6 carries the wedge γ as a free parameter over $[0, 0.03]$ per kWh rather than taking a side. Figure D.2 shows the level conclusion survives the whole range, because carbon dominates; what γ governs is the distributional character of the status quo, which is the subject of Section 6.6.

Two features make this externality unusual among intermediate inputs. It is *badly matched to the value of the good*. Where abatement is possible at the point of harm, a tax on the input is a poor instrument for any externality, which is the standard targeting principle. What is unusual here is the *dispersion* of the harm-to-value ratio. Energy per query varies by more than two orders of magnitude across workloads, from a short text completion to an agentic task, while price varies far less, so the ratio of external cost to price differs several-fold across the queries a single ad valorem rate would cover. While two distinct tons of steel are likely similar in their energy costs, the same is not necessarily true of two queries. Section 6 quantifies this.

It is also *spatially discretionary in the extreme*. A steel mill's emissions are tied to where the mill is, and the mill is where the ore and the workforce are. An inference request can be served from anywhere with power and fibre, at a latency cost measured in milliseconds. The elasticity of the location of the harm with respect to a local charge is therefore far higher than for any conventional polluting activity, which is what makes leakage a first-order concern and motivates the destination-based backstop of Appendix B.3.

Market power in supply. Frontier model provision is a natural oligopoly. Training a frontier model costs hundreds of millions to billions of dollars and the cluster required is comparable; marginal cost per token is fractions of a cent. Reported gross margins on frontier inference run to fifty or sixty percent, implying a markup around 2.5 against an economy-wide 1.25 (De Loecker et al., 2020). A good sold above marginal cost is consumed too little, which is the claim that runs opposite to the other three. Section 5.6 shows how much of that markup is genuine rent rather than fixed-cost recovery, and how much of the resulting correction survives free entry.

Foreign ownership of the rents. Those rents accrue disproportionately to shareholders in a small number of countries. For a jurisdiction that imports AI services, a tax on usage is partly a tariff that shifts rent across the border, and the optimal-tax calculus differs accordingly. This is why ownership rather than technology turns out to set the sign of the optimal wedge.

None of these features is unique to AI taken singly. Electricity has unpriced carbon, pharmaceuticals have market power, and imported goods carry terms-of-trade considerations. What is unusual is their coincidence in one input. That is why compute clears a bar that a lathe, a truck or a ton of steel does not, and why the four arguments have to be priced against each other rather than in turn.

3 Model

Table D.4 collects every parameter of the model with its calibrated value, range and source, and serves as a complete overview of its primitives.

3.1 Environment

Tasks and production. A single final good Y is produced competitively from two task bundles. A share $\lambda \in (0,1)$ of production value comes from *automatable* tasks, in which AI and routine labor are substitutes. The remaining share comes from *non-automatable* tasks, performed by non-routine labor. With X_A and X_M the two bundles,

$$Y \;=\; \Lambda\, X_A^{\lambda}\, X_M^{1-\lambda}, \qquad X_A \;=\; \Big[\beta\, L_R^{\frac{\sigma-1}{\sigma}} + (1-\beta)\, a^{\frac{\sigma-1}{\sigma}}\Big]^{\frac{\sigma}{\sigma-1}}, \qquad X_M \;=\; L_N, \tag{1}$$

where a is AI services, L_R and L_N are routine and non-routine labor, and Λ is a constant.[7]

Write q for the buyer price of AI relative to the routine wage, c_A for the unit cost of the automatable bundle, and

$$s \;\equiv\; \frac{q\,a}{c_A X_A} \;=\; \frac{(1-\beta)^{\sigma} q^{1-\sigma}}{\beta^{\sigma} + (1-\beta)^{\sigma} q^{1-\sigma}} \qquad \text{so that} \qquad \frac{\mathrm{d}\ln s}{\mathrm{d}\ln q} \;=\; -(1-s)(\sigma-1), \tag{2}$$

for the AI cost share within that bundle. Exposure, the share of national income spent on AI, is $\theta \equiv \lambda s$.

[7]The Cobb–Douglas outer aggregator is for transparency: it fixes the value shares of the two bundles and lets every result be stated in cost shares. A CES outer nest changes magnitudes but not structure; Lemma D.2 bounds the correction.

The elasticity σ is the central technological parameter: values above one make AI a gross substitute for routine labor within the automatable bundle, which is what gives the displacement concern its content, and $\sigma = 1$ shuts the redistributive channel entirely. It is also the parameter we know least well, so Appendix D.1 reports every result across $\sigma \in [1.2, 3]$. The outer Cobb–Douglas nest is not innocuous either, and Lemma D.2 in Appendix A records what changes when it is relaxed.

The AI sector. AI services are supplied by a sector with market power. One unit requires compute at resource cost m_0 and e kilowatt-hours of electricity at the price p_e paid by data centers. Providers also choose an efficiency level $\xi \geq 0$, which stands for chip generation, model architecture, serving stack and cooling. It lowers energy per unit but not the carbon intensity of the grid serving the site, ζ, which we take as given; cleaner siting is priced through site-specific rates rather than through ξ. Efficiency lowers energy per unit, $e = e(\xi)$ with $e' < 0$ and $e'' > 0$, at a resource cost $\varphi(\xi)$ with $\varphi' > 0$ and $\varphi'' > 0$. Marginal cost and the efficiency choice are

$$m(\xi) \;=\; m_0 + \varphi(\xi) + p_e\, e(\xi), \qquad\qquad \varphi'(\xi) \;=\; -\,p_e\, e'(\xi). \tag{3}$$

The second condition is the engine of the targeting result below. The efficiency margin responds to the price of energy and to nothing else. It does not respond to the price of AI output.

Providers are an oligopoly: n symmetric firms with the constant marginal cost $m(\xi)$ of (3) and a fixed cost F covering the training run and the compute cluster. They compete in quantities, so the equilibrium buyer price carries a markup over marginal cost,

$$p \;=\; \mu\, m(\xi)\,(1 + t_U), \qquad \mu \;=\; \frac{\sigma_q}{\sigma_q - 1/n} \;>\; 1, \tag{4}$$

decreasing in n. Entry is free, so n adjusts until profit is zero; Section 5.5 shows this makes μ a function of the tax rather than a parameter, which is where the entry amplification comes from.

Because the same sector must deliver both the targeting result and the entry result, we fix one market structure and use it throughout rather than switching between a monopolist for one argument and free entry for another.[8]

[8] Appendix C.4 records what this buys and what it costs.

Externalities. Electricity for AI produces carbon at intensity ζ kg per kWh, valued at social cost δ, and its delivery causes capacity and transmission costs of which a portion γ per kWh is recovered from other ratepayers rather than from the facility. Write $d \equiv \zeta(\delta - \tau_c) + \gamma$ for the residual external cost per kilowatt-hour, where τ_c is any carbon price already in force, and $D \equiv d\, e(\xi)$ for the external cost per unit of AI. Two features of d matter. It is defined per kilowatt-hour, not per unit of output, so the mapping to a usage tax runs through $e(\xi)$ and therefore through an endogenous object.

The second feature is the hinge of the targeting argument: the externality can be reduced either by using less AI or by making each unit cleaner, and only an instrument that reaches the second margin can induce efficiency investment. Appendix B.8 sets out both margins.

Households: types are ex ante, tasks are chosen. We depart here from the two-type structure used in most of this literature, and the departure changes the object welfare weights attach to. If workers are permanently assigned to tasks, "routine worker" is a type and a welfare weight can be attached to it. If workers sort across tasks in response to wages[9] then "routine worker" is an *outcome*, its membership moves when the tax moves, and a welfare weight attached to it is not a well-defined object. We therefore put the sorting margin in the primitive model and define welfare on the objects that do not move.

There is a unit mass of workers, divided into two *ex ante* groups $i \in \{1, 2\}$ of equal mass. A worker of group i draws task-specific efficiencies (z_R, z_N) independently from a Fréchet distribution with shape $\kappa > 0$ and group-specific scales A_{iR}, A_{iN}, and supplies her labor to the task with the higher return, so the wage she commands is $\max_j A_{ij} z_j w_j$. Groups differ only in comparative advantage,

$$\varrho \;\equiv\; \ln \frac{A_{1R}}{A_{1N}} \;-\; \ln \frac{A_{2R}}{A_{2N}} \;>\; 0, \tag{5}$$

so group 1 is the group with comparative advantage in the tasks AI reaches. Group membership is the primitive: it is fixed, unobservable, and untaggable. Occupation is not. Hours are supplied with elasticity ε_L.

Two objects follow from the Fréchet assumption and carry the whole distributional analysis: the share of each ex ante group working automatable tasks, and the elasticity of that share to the relative wage. Their closed forms, and the exposure gap Δ_ℓ and reallocation elasticity $\mathcal{E}$ built from them, are in Appendix B.8.

[9] This is the empirically relevant case, and the one Section 5.4 takes seriously.

Government and instruments. The government levies a linear income tax τ with a lump sum transfer and cannot condition either on group. It also has access to a usage tax t_U, a cash flow tax z on the supernormal profits of providers, and a Pigouvian charge t_E per kilowatt-hour at the data center. Social welfare is Bergson–Samuelson with weights g_i defined on *ex ante groups*, and we write $\hat{g}_1 \equiv g_1 - \bar{g}$ for the excess weight on group 1 relative to the income-weighted average, normalized so that a uniform proportional change in all wages has no first-order distributional value. Because the weights sit on groups rather than on occupations, they are invariant to the reform, and the normalization $\sum_i \hat{g}_i W_i = 0$ holds before and after.

We assume throughout that τ is set optimally. This permits the envelope argument of Lemma 3: a reform that raises all wages proportionally has no first-order welfare effect, because the income tax has already been adjusted to the point where the marginal value of redistribution equals its marginal cost. Only the *relative* wage movement survives into the optimal-tax formula, which is what makes the predistribution term a statement about the wage *gap* rather than about the wage level.

Assumption 1 (Restricted instruments)**.** The government cannot tag workers by ex ante group, cannot observe the AI intensity of individual jobs, and is restricted to a linear income tax, so it cannot implement a schedule that separates groups perfectly.

3.2 Equilibrium and comparative statics

All proofs are collected in Appendix A.

Factor payments follow from (1). The automatable bundle receives λY, of which a share $1-s$ goes to routine labor and s to AI; the non-automatable bundle receives $(1-\lambda)Y$, all of it to non-routine labor:

$$w_R L_R = \lambda(1-s)\,Y, \qquad w_N L_N = (1-\lambda)\,Y, \qquad q\,a = \lambda s\,Y = \theta Y. \tag{6}$$

Lemma 1 (Exposure)**.** *A small increase in the buyer price of AI raises the unit cost of the final good, at given factor prices, at a rate equal to the AI cost share:* $\mathrm{d}\ln c_Y / \mathrm{d}\ln q = \theta = \lambda s$*. Equivalently, real income evaluated at given factor prices falls by* $\theta\,\mathrm{d}\ln q$*.*

Lemma 1 is a money-metric statement: it gives the real-income cost of a price change before labor and AI quantities adjust, not the equilibrium response of gross output, which also depends on those adjustments. It imposes a useful discipline. Because exposure is small, both the real-income cost of a usage tax and its revenue are small. Whatever the case for such a tax, it is not a fiscal case, and Section 6 puts numbers on this.

Lemma 2 (Relative wages and the mobility elasticity)**.** *Let $\mathcal{E}$ denote the elasticity of relative task labor supply to the relative task wage generated by reallocation. Then*

$$\frac{\mathrm{d}\ln(w_R/w_N)}{\mathrm{d}\ln q} \;=\; \frac{s\,(\sigma-1)}{1+\varepsilon_L+\mathcal{E}}, \qquad \textit{and under (B.3)} \qquad \mathcal{E} \;=\; (\kappa-1)\Big[1-\frac{\pi(1-\pi)\,\Delta_\ell^2}{h(1-h)}\Big],$$

where π is group 1's share of employment and h the routine share. In the symmetric case $\pi = h = \frac{1}{2}$ this reduces to $\mathcal{E} = (\kappa-1)(1-\Delta_\ell^2)$. A higher price of AI compresses the task wage gap if and only if AI and routine labor are gross substitutes, and the compression is damped by hours adjustment and by reallocation.

Reallocation is measured in efficiency units, because production uses effective task labor rather than headcount: under Fréchet sorting the workers who switch are less productive in their new task than the incumbents, so effective supply responds with elasticity $\kappa-1$ rather than κ, and $\kappa > 1$ is required for the expected wage in (B.3) to be finite. The identity is the hinge of everything that follows. The Fréchet shape κ is not "mobility": it governs the dispersion of comparative advantage, and so determines both how sharply groups sort into tasks and how much probability mass sits near the switching margin. These pull opposite ways, and Section 5.4 works out the consequence, which is that the fixed-assignment two-type model is the limit $\kappa \to \infty$ rather than $\kappa \to 0$.

Corollary 1 (The labor share)**.** *Aggregate labor income is $(1-\theta)Y$, and $\mathrm{d}\ln(1-\theta)/\mathrm{d}\ln q = \lambda s(\sigma-1)(1-s)/(1-\lambda s) > 0$ if and only if $\sigma > 1$. Cheaper AI lowers the labor share when AI and routine labor are gross substitutes.*

Finally, the own price elasticity of AI demand combines the within-bundle substitution effect with the scale effect of Lemma 1:

$$\sigma_q \;\equiv\; -\frac{\mathrm{d}\ln a}{\mathrm{d}\ln q} \;=\; \underbrace{\sigma(1-s)}_{\text{substitution}} \;+\; \underbrace{s+\theta}_{\text{scale}} \;\approx\; \sigma \quad \text{when } s,\theta \text{ are small.} \tag{7}$$

Two results are needed before the tax formula, because each underpins a step that is otherwise a leap. The first makes precise the sense in which the income tax neutralizes the common wage movement. The claim is often asserted; here we state exactly what it requires.

Lemma 3 (Envelope neutrality of the common wage component)**.** *Suppose the government sets the linear income tax (τ, T) optimally, worker preferences are weakly separable between consumption and labor, and the wage response to a reform of t_U decomposes into*

a common proportional component $\mathrm{d}\ln\bar{w}$ *and a relative component* $\mathrm{d}\ln(w_R/w_N)$*. Then the common component carries no first-order* distributional *value. At the optimal linear tax, the private value of a common income change, weighted by the mean welfare weight, plus its fiscal externality through labor supply and the income tax base, sums to its value in units of public funds. All welfare weights in the paper's formulas are effective weights in those units, so that an average private dollar is worth one, and the distributional value of the reform is carried entirely by the relative component, weighted by the excess weight* $\hat{g}_1$*. The decomposition is exact when the wage change is spanned by a common and a relative component, as in the two-task structure of the model; departures from it, such as heterogeneous exposure within a group, are not quantified here.*

The decomposition into a common and a relative part is a property of the technology rather than an assumption: with constant returns and competitive factor markets each wage equals its marginal product, and a Cobb–Douglas or CES task structure makes any shock resolve into a term that scales all marginal products together and a term that tilts them. The same structure underlies the general-equilibrium incidence analysis of Sachs et al. (2020), and the neutrality of the common component under an optimal linear tax is the Naito–Saez logic (Naito, 1999; Saez, 2004), itself a linear-tax specialization of Stiglitz (1982). The shortcut is not free: it holds exactly when the income tax spans the distributional content of the common wage movement, which for a linear tax means a proportional shift that τ and T can undo. A technology that fattens one tail of the earnings distribution violates this, and the residual term the lemma isolates is what a richer nonlinear analysis would carry. We flag at each use that the shortcut rests on Lemma 3 and is exact only under its covariance condition.

The cancellation of exposure does not depend on the Cobb–Douglas outer nest, but the size of the wage response does. Lemma D.2 in Appendix D.10 replaces it with a CES aggregator of elasticity η between the automatable and non-automatable bundles and shows that the predistribution term scales with $(\sigma-\eta)/(\eta+\varepsilon_L+\mathcal{E})$ rather than $(\sigma-1)/(1+\varepsilon_L+\mathcal{E})$. What matters is how substitutable AI is for routine labor *relative to* how substitutable the two bundles are, and at $\sigma=\eta$ the channel vanishes. Over $\eta\in[0.5,1.5]$ the term ranges from about 0.4 to 2.1 times its Cobb–Douglas value at the baseline. Since η is not well measured for this split, this widens the uncertainty around the predistribution term considerably; it does not change the sign of the domestic wedge, which the market-power term governs.[10]

[10]We use Cobb–Douglas in the body for readability.

3.3 The benchmark

Proposition 1 (Production efficiency)**.** *Suppose there are no externalities ($d = 0$), the AI sector is competitive, and the government can tag worker types, so that Assumption 1 fails. Then $t_U^\star = 0$. Any revenue raised by a tax on AI purchases can be raised at strictly lower welfare cost by income and final good taxes.*

Proposition 1 is the right starting point and it is frequently forgotten. A usage tax distorts the choice of how to perform a task, raises the cost of every unit of output that uses AI, collects revenue on a base that is currently trivial, and buys nothing that the income tax cannot buy more cheaply. Every argument for such a tax must therefore be an argument that one of its hypotheses fails. There are exactly three ways for that to happen here, and the next section prices each of them.

One hypothesis deserves naming rather than leaving implicit, because the entire predistribution term rests on it. Assumption 1 restricts the government to a *linear* income tax, and that restriction is what makes the general-equilibrium wage response something the planner wants to manipulate—the Naito channel. It is not harmless. Scheuer and Werning (2016) show that in exactly the assignment and Roy models we use, general-equilibrium wage effects wash out of the optimal *nonlinear* income tax, which reverts to its fixed-wage Mirrlees form, and production efficiency is restored. Read through that result, our predistribution term measures not a feature of the technology but how far the income tax falls short of the nonlinear optimum. We keep the linear restriction because it describes most real tax systems' capacity to target occupational exposure, and because it is the assumption most favorable to the instrument we argue against. But the term belongs in the right category: an upper bound generated by a restricted instrument set.

4 The optimal tax on AI usage

Consider a marginal reform dt_U, with the income tax adjusted to balance the budget and set optimally, and with the number of providers held fixed; Section 5.6 turns to free entry. Welfare in money metric moves for four reasons. Firms adjust their AI use, and each unit they no longer buy had a private value q but a social resource cost m. With the number of providers fixed, the fixed cost of provision is sunk and does not vary with output, so the entire margin of price over marginal cost separates private from social value on this margin. Each unit also carried an external cost D. Relative wages move, which has distributional value because types cannot be tagged. All purely mechanical transfers

cancel, and by Lemma 3 the common component of the wage response has no first-order distributional effect.[11]

Proposition 2 (The optimal usage tax)**.** *Under Assumption 1, with the income tax set optimally, the number of providers fixed and constant-elasticity demand for AI, the optimal ad valorem tax on AI purchases satisfies*

$$\boxed{t_U^\star \;=\; \underbrace{\frac{(1-\tau)\,\hat{g}_1\,\Delta_\ell\,(\sigma-1)\,\beta_1}{(1+\varepsilon_L+\mathcal{E})\,\sigma_q}}_{\text{predistribution, } \Omega/p} \;+\; \underbrace{\frac{D}{p}}_{\text{Pigouvian}} \;-\; \underbrace{(1-f)\left(1-\frac{1}{\mu}\right)}_{\text{market power}}} \tag{8}$$

where $\beta_1 \equiv B_1/(\lambda Y)$ is the gross earnings of ex ante group 1 relative to the value of the automatable bundle, Δ_ℓ is the exposure gap of (B.4)*, and $\mathcal{E}$ the reallocation elasticity of Lemma 2. Under fixed assignment group 1 is the routine workforce, so $\beta_1 = 1 - s$ and the first term reduces to the two-type formula at $\Delta_\ell = 1$, $\mathcal{E} = 0$. When s and θ are small, $\sigma_q \to \sigma$.*

Each term answers a different question in the debate.

The predistribution term. This is the robot tax. It is proportional to the excess welfare weight $\hat{g}_1$ on the group the planner wants to help, which is a normative parameter. It is proportional to the exposure gap Δ_ℓ, which measures how much of a relative *task* wage movement survives as a relative movement between the groups the planner cares about, and which is a measurable feature of the sorting pattern rather than a modelling convention. It is proportional to the earnings base β_1 of the group the planner weights: once workers sort, that group earns income in non-routine tasks too, so its earnings exceed routine earnings whenever $\Delta_\ell < 1$. And it is proportional to $(\sigma - 1)$, so it vanishes at $\sigma = 1$, since a change in the price of AI then leaves the routine wage share of the automatable bundle unchanged. It is damped by the labor supply elasticity and the reallocation elasticity, because workers then absorb part of the shock in hours and in switching rather than in wages, and by the net of tax rate, because the income tax already absorbs part of any wage movement.

The literature almost universally sets $\Delta_\ell = 1$ by identifying the losing group with the occupation. That is not a conservative simplification but the corner of the parameter space that maximizes the term, and it asserts that every member of the disadvantaged

[11] Throughout, q is the buyer price relative to the routine wage. A usage tax moves q one for one at given wages; the induced wage response feeds back on q only through terms of order s, which we neglect.

group works in an exposed task and no member of the advantaged group does. Section 5.4 shows what replaces it and Section 6 shows that the exposure gap is the tightest of the three factors.

The Pigouvian term. The corrective component is the residual external cost as a share of the pre-tax price, exactly as in Sandmo (1975). It is not the external cost as a share of the electricity price, and that distinction does the analytical work in Section 6. It also depends on the provider's efficiency choice through $D = d \cdot e(\xi)$, and Proposition 3 shows that the usage tax cannot move ξ.

The market power term. This term is negative, and at any markup above about ten percent it is the largest of the three, so most of the argument about the sign of the optimal tax is an argument about it. With the number of providers fixed, the fixed cost of training and capacity is sunk: it does not change when output changes, so the whole margin $1 - 1/\mu$ of price over marginal cost is underprovision. Recovering a fixed cost does not remove the distortion, since it is the reason the price exceeds marginal cost in the first place. Two qualifications matter and are developed in Section 5.6. The term values provider profit, so its proper weight is the planner's weight on shareholders rather than one. And it relies on constant-elasticity demand, under which the producer price does not respond to the tax; when providers absorb part of a tax, the tax also extracts rent, and Section 6.5 shows that this can reverse the sign.

Why does ownership matter? If the problem is that a concentrated sector produces too little, ownership should seem irrelevant to the remedy. With the number of providers fixed it is not. The producer price does not move, so a marginal tax change falls entirely on buyers, whose loss on the units they keep buying is exactly the revenue the government collects. What remains is the margin on the units no longer sold, and that margin is provider profit. Profit accruing abroad is not part of domestic welfare. This conclusion is specific to blocked entry: under free entry profit is zero, the gain from lower prices accrues to buyers, and ownership drops out (Proposition 8).

Corollary 2 (Ownership and the sign of the tax, with entry blocked)**.** $\partial t_U^\star / \partial f = 1 - 1/\mu \geq 0$. *An economy that imports all of its AI from foreign providers ($f = 1$) faces $t_U^\star = \Omega/p + D/p$, which is positive whenever the exposure gap is positive and emissions are not already priced at source, while an economy that owns its providers faces a strictly lower rate, negative whenever the Lerner index $1 - 1/\mu$ exceeds $\Omega/p + D/p$.*

Corollary 2 is consistent with an observed asymmetry: jurisdictions hosting the frontier labs are hostile to taxing AI usage, while jurisdictions importing AI services are legislating digital levies. The importer's positive rate is not a terms-of-trade gain, since with constant-elasticity demand the foreign producer price does not fall. It arises because the domestic planner does not count the foreign profit that underprovision would otherwise protect. Two caveats bound it. It holds only while entry is blocked, and it is the most sensitive of the paper's results to how much of a tax providers absorb (Section 6.5).

Remark 1 (Tax base erosion is not an argument). It is often argued that AI must be taxed because it shrinks the labor income tax base. Corollary 1 confirms the premise and Proposition 2 rejects the inference. The mechanical revenue loss from a smaller labor share is not a first-order welfare loss: the government raises τ, at a marginal cost of public funds already reflected in the optimal τ. A tax base is not a policy objective. What matters is who is made worse off, and that is what the term in $\hat{g}_1$ measures. The base erosion argument is a distributional argument wearing a fiscal costume.

5 Reading the three terms

Proposition 2 gives three terms and says nothing about how large they are or why one of them is negative. This section takes them in turn. The organizing idea is instrument assignment: each of the three problems the debate raises has an instrument built for it, and the question for the usage tax is what remains once those instruments are doing their work. Sections 5.1 to 5.3 make the assignment. Sections 5.4 and 5.6 then derive the two results that shrink the terms most—the exposure gap on the redistributive side, and the ownership weight and entry offset on the market-power side—and explain the negative sign that the last of them produces.

Proposition 2 describes the second best when the usage tax is the only instrument. It is not. This section shows that each term in (8) has a better instrument, then develops three extensions that the static formula omits. They do not all push the same way: the exposure correction lowers the predistribution term while the earnings base raises it, a shareholder weight below one shrinks the negative market-power term, and free entry replaces that term with one that does not depend on ownership.

5.1 Emissions belong on the emissions node

The general principle is Kaplow (2012): when an income tax is available, the corrective instrument should be set at the Pigouvian level on the externality itself and the distributional consequences handled by the tax and transfer system, rather than folded into the corrective rate. What the following proposition adds is that in this setting there are two externality margins, not one, and only one instrument reaches both.

Proposition 3 (Targeting, with an efficiency margin)**.** *Let t_E be a charge per kilowatt-hour at the data center. Setting $t_E = d$ implements the first best on both the scale and the efficiency margins and drives the Pigouvian term in* (8) *to zero. A usage tax set to D/p replicates the first best on the scale margin only. The welfare gap per unit of AI is*

$$\Delta \;=\; \Big[\varphi(\xi^0) + (p_e + d)\, e(\xi^0)\Big] \;-\; \Big[\varphi(\xi^\star) + (p_e + d)\, e(\xi^\star)\Big] \;\geq\; 0,$$

where ξ^0 denotes the efficiency level satisfying the second condition in (3)*, the first-order condition $\varphi'(\xi) = -p_e\, e'(\xi)$, evaluated at the energy price p_e, and $\xi^\star$ the level satisfying that same condition when the perceived energy price is $p_e + d$. The gap is strictly positive whenever any efficiency response is technically possible.*

The efficiency margin is not a curiosity. It is where most of the action has been. Energy per query has fallen by roughly an order of magnitude in three years through better chips, better serving stacks, and smaller distilled models, and the carbon intensity of a query varies by a factor of several across grids. A tax on tokens is neutral with respect to all of it.

Corollary 3 (Location)**.** *With heterogeneous grid intensities ζ_i across sites, a uniform usage tax applies the same wedge to a query served from a clean grid and one served from a coal grid. An emissions charge applies wedges in the ratio ζ_i/ζ_j. The usage tax cannot implement the efficient allocation of load across sites at any positive rate.*

A note on tax form. Everything above concerns the tax *base*—energy versus the value of AI output—and is separate from the question of tax *form*, ad valorem versus specific. The two interact, because under imperfect competition the choice of form affects pass-through and the equilibrium markup. Appendix B.1 shows that the targeting conclusion is unaffected: the choice between a specific and an ad valorem levy matters, and under imperfect competition the ad valorem form has the edge, but both are levied on a base that is the wrong physical quantity, and no choice of form repairs a base that mistracks the harm by an order of magnitude.

5.2 Rents belong on the rent

Proposition 4 (Neutrality of the rent tax)**.** *A cash flow tax at rate z on supernormal profits, that is on economic profit net of the normal return to capital, with immediate expensing and full loss offset, leaves the provider's first order conditions for price, quantity, and efficiency unchanged. It raises $z\Pi$ with zero first-order deadweight loss, so as a source of revenue it dominates a usage tax raising the same amount whenever $\sigma_q > 0$. A usage tax may still be warranted for the corrective reasons of Proposition 2, which a rent tax does not address.*

Proposition 4 does not reduce the markup and so does not remove the negative term in (8). It transfers rent to the treasury without changing the allocation, which is a different question from the one Proposition 2 answers. Proposition 2 asks what wedge belongs on the *transaction*, and finds that genuine rent argues for a lower one; Proposition 4 asks how to collect the rent itself, and finds an instrument that collects it without moving any wedge at all. The two are complements: the rent tax raises revenue from the same rent whose existence made the usage tax look worse.

The tax logic here is standard. Cash flow bases with expensing and full loss offset are the standard neutral instrument (Auerbach and Devereux, 2018), and the contribution here is the mapping to this sector: the base is the excess of revenue over expensed cost, the sunk compute stock cannot flee, and the elasticity that makes the usage tax expensive is absent by construction. The known failure mode is political rather than technical. What gets legislated is typically a levy on gross revenue—a digital services tax—which falls on turnover rather than cash flow and so lands on marginal as well as inframarginal units. A revenue tax and a cash flow tax are not variants of one instrument but opposites: one distorts the quantity margin exactly as a usage tax does, the other does not.

One caveat matters in practice: neutrality requires that the base be economic rent and not the normal return.[12]

5.3 Redistribution belongs on the income tax

Corollary 4 (When predistribution is worth nothing)**.** *If the government can implement a nonlinear income tax that separates groups, or can tag at low cost on a variable that identifies group membership, then at the optimum $\hat{g}_1 \to 0$ and the predistribution term*

[12] A tax on gross revenue, which is what most digital services taxes actually are, is a usage tax in disguise and inherits every one of its distortions. The distinction between a revenue tax and a cash flow tax is the distinction between Proposition 2 and Proposition 4, and it is not a detail.

vanishes. A tag that is only correlated with membership leaves heterogeneity within cells, and so shrinks $\hat{g}_1$ without eliminating it. A related result holds without tagging if the income tax is unrestricted (Scheuer and Werning, 2017), although there the wage channel loses its policy value rather than the group weights becoming equal.

The predistribution term is a statement about missing instruments, not about technology, and there are three independent routes to zero. Tag, or use a nonlinear schedule, and $\hat{g}_1$ collapses. Without tagging, Scheuer and Werning (2016) show that in a Roy economy the general-equilibrium wage channel is not a lever a fully optimizing income tax wants to pull, because the schedule already prices every earnings level at which the wage change lands. The alternative option is to wait: Guerreiro et al. (2022), in a model with routine and non-routine occupational choice close to ours, find a positive robot tax survives only while the cohort that cannot retrain is still working. Any AI tax justified on this basis is transitional by construction and should be legislated with a sunset.

Kleven and Zidar (2026) give the cleanest statement of the limit. When labor supply responds only on the extensive margin and each occupation pays a distinct wage, a nonlinear earnings tax is occupation-specific, and any redistribution that equilibrium wages perform the tax system can perform directly; the predistribution term is then exactly zero. Their neutrality fails once workers of different types can earn the same income, which is the case in the Roy economy of this paper: both groups work in both tasks, so earnings do not reveal the group, and the term survives in proportion to the exposure gap. The empirical question is therefore how far the economy is from their benchmark, and Table 2 shows that the surviving term is small and, on the direct AI measures, of the opposite sign.

5.4 The predistribution term: mobility and identification are one parameter

Section 3 put the sorting margin in the model rather than bolting it on, and this subsection collects what that buys. The question it answers is the one the robot-tax literature usually settles by assumption: how much of the predistribution term survives when workers can move?

The intuitive answer is that mobility damps the term. A fall in the routine wage induces exit from routine tasks, which raises the routine wage back toward where it started; in the limit of frictionless reallocation, relative task wages cannot be moved at all by an input tax and the term is zero. That argument is correct about *task wages* and it

is the reason $\mathcal{E}$ appears in the denominator of (8). It is incomplete as a statement about welfare, and the incompleteness reverses a comparative static.

Proposition 5 (Mobility and identification are jointly constrained)**.** *Under the Roy structure of* (B.3)*, in the symmetric case of Lemma 2, where group earnings and task earnings coincide, the predistribution term is*

$$\frac{\Omega}{p} \;=\; \frac{(1-\tau)\,\hat{g}_1\,\Delta_\ell\,(\sigma-1)(1-s)}{\big(1+\varepsilon_L+(\kappa-1)(1-\Delta_\ell^2)\big)\,\sigma_q}, \qquad \Delta_\ell \;=\; \tanh\!\big(\tfrac{\kappa\varrho}{4}\big).$$

Then, for $\kappa > 1$ and holding the comparative-advantage gap ϱ fixed:

(i) the reallocation elasticity $\mathcal{E} = (\kappa-1)(1-\Delta_\ell^2)$ rises from zero and returns to it, with $\mathcal{E} \to 0$ as $\kappa \to 1$ and as $\kappa \to \infty$;

(ii) the exposure gap Δ_ℓ is strictly increasing in κ, from $\tanh(\varrho/4)$ to 1;

(iii) consequently Ω/p is strictly increasing *in κ, rising to the fixed-assignment value as $\kappa \to \infty$, and this holds for every $\varrho > 0$ without further restriction. The two-type model of the robot-tax literature is the $\kappa \to \infty$ limit;*

(iv) for any fixed measured exposure gap $\Delta_\ell < 1$, the term is bounded above by Δ_ℓ times its fixed-assignment value, with equality only where effective reallocation vanishes, at $\kappa = 1$.

Parts (i) to (iii) invert the comparative static usually reported in this literature. Treating the Fréchet shape as a mobility parameter, so that raising it damps the predistribution term while the welfare weight on "routine workers" stays fixed, is not a valid exercise: it varies the reallocation response while holding constant an object—the identity of the losing group—that the same parameter determines. The inversion is a statement at a given comparative-advantage gap ϱ. If ϱ is allowed to adjust so that the exposure gap stays fixed, higher κ does raise reallocation and lower the term; what the model rules out is treating κ as a pure mobility parameter while ϱ and the identity of the losing group are both held fixed. Once both move together, the sign flips. Perfect sorting is perfect segregation. A world in which comparative advantage is sharp enough for the exposed group to be cleanly identifiable is also a world in which nobody is near the switching margin, and the fixed-assignment model is recovered exactly.

Part (iv) is where the economics is. The right way to discipline the term is not to argue about κ but to measure the two objects the formula contains. The exposure gap Δ_ℓ

is estimable from any occupational exposure measure crossed with any ex ante grouping, and $\mathcal{E}$ is then pinned by Lemma 2. The literature's convention $\Delta_\ell = 1$ is not a conservative simplification. It is the corner of the parameter space that maximizes the term, and it is empirically false. Section 6 measures it and finds the term falls to about half of the fixed-assignment number: the exposure gap cuts it by more than half and reallocation trims it further, while the broader earnings base of the group the planner weights restores part of the loss.

Two implications for policy survive intact from the earlier treatment, and one does not. What survives: the term is transitional, because ϱ itself falls as cohorts choose skills knowing what AI can do, which is the parameter behind the sunset that Corollary 4 calls for.[13] What does not survive is the claim that a reader who believes in fast reallocation should read the predistribution column as zero. The right statement is that a reader who believes the losing group is sharply identifiable should also believe reallocation is slow, and vice versa, and that neither belief rescues the term, because the product is bounded by the exposure gap either way. Figure B.2 draws both halves of the proposition; Appendix B.10 develops the frontier, the horizon dependence of ϱ, and two caveats that cut in opposite directions.

5.5 Endogenous entry: the usage tax amplifies market power

Taxation with an endogenous entry margin is studied by Scheuer (2014) in the context of entrepreneurship, where the same tension appears: an instrument that looks like it falls on rents partly falls on the decision to be in the market at all.

The markup was taken as given above. It is not. Provision has large fixed costs in research and in compute clusters, so the number of providers is determined by free entry, and a tax that shrinks the market shrinks the number of firms.

Let n symmetric providers compete in quantities facing the demand elasticity σ_q of (7), so the equilibrium markup is $\mu = \big(1 - 1/(n\sigma_q)\big)^{-1}$, and let each incur a fixed cost F. Free entry drives per-firm profit to zero.

The result below is the one that most surprises people, so the mechanism is worth stating before the statement. A tax on usage raises the price buyers pay and shrinks the market. A smaller market supports fewer firms, because each must still cover the same fixed cost. Fewer firms compete less hard. So the tax, which was meant to fall on incumbents, ends by handing the survivors a wider margin on the sales they retain. The

[13] The correction runs in the same direction as every other extension in this section, lowering the optimal usage tax.

tax is a barrier to entry that the incumbents did not have to build.

Proposition 6 (Entry amplification)**.** *With Cournot competition, constant-elasticity demand and free entry, on the regular branch $\mu < (\sigma_q + 1)/(\sigma_q - 1)$, a usage tax raises the equilibrium markup,*

$$\frac{\mathrm{d}\ln\mu}{\mathrm{d}\ln(1+t_U)} \;=\; \frac{\sigma_q(\mu-1)}{(\sigma_q+1)-(\sigma_q-1)\mu} \;>\; 0,$$

and the buyer price responds by

$$\frac{\mathrm{d}\ln q}{\mathrm{d}\ln(1+t_U)} \;=\; \Phi \;\equiv\; \frac{\mu+1}{2-(\mu-1)(\sigma_q-1)} \;>\; 1.$$

An emissions charge that raises marginal cost also induces exit when $\sigma_q > 1$, with buyer-price elasticity $\Phi_E = [1 - \frac{1}{2}(\mu-1)(\sigma_q-1)]^{-1}$ with respect to marginal cost.

The amplification is large. At $\mu = 1.5$ and $\sigma_q = 1.965$, $\Phi = 1.65$: a usage tax of one percent raises the buyer price of AI by about 1.65 percent, because it also drives out entrants and lets the survivors raise their margins; at $\mu = 2$, $\Phi = 2.9$. The expression solves the price and entry conditions jointly, and solving the free-entry equilibrium numerically reproduces it at every markup we have tried.[14] Entry is not specific to the usage tax: the emissions charge also shrinks producer revenue and induces exit, though less strongly, since it enters the entry condition through marginal cost rather than through the full tax factor. What the amplification implies for the optimal tax depends on the welfare accounting under free entry, which Section 5.6 develops.

Training compute. A token tax is levied on queries served, so it misses the training that produced the model. Training draws electricity and emits carbon but produces no tokens, so its implicit rate under a token tax is not low but exactly zero, which leaves a substantial share of the sector's electricity outside the base, from a tenth to two thirds depending on which estimate of the inference share one takes. A charge at the meter reaches it automatically, because a kilowatt-hour used for training is billed like one used for a query. If training also creates risk that rises more than proportionally with compute, the corrective instrument is a levy on training compute that tracks marginal risk, which

[14]The Cournot structure is doing work here: with contestable markets or a competitive fringe $\Phi \to 1$, and with linear demand the same structure gives $\Phi < 1$. What survives under the demand system used here is the direction: with constant-elasticity demand, large fixed costs and entry that responds to market size deliver $\Phi > 1$.

can improve on the compute thresholds in current regulation, since a threshold is a tax rate of zero followed by one of infinity (Proposition B.2). Either way, the base should be electricity or compute, not tokens.

The gap grows over time. Proposition 3 compares instruments at a fixed efficiency technology, and in doing so understates its own conclusion. Efficiency is the outcome of durable investment, and only an instrument that raises the perceived price of energy enters the investment condition at all. Appendix B.5 extends the model to a horizon T and shows that the emissions charge induces strictly more efficiency investment at every horizon and that the cumulative welfare gap between the instruments is increasing in T; under an isoelastic condition the efficiency gap itself also widens. This argues for getting the base right early, because durable investment responds to the base that is in place.

5.6 The market-power term in two regimes

The market-power term is the largest object in the formula and the only one that is negative at every markup, so it governs the sign of everything reported later. It is also the term hardest to accept at face value, since it appears to say that a planner should subsidize the most profitable firms in the economy. This section derives it from a single welfare calculation that tracks buyers, provider cash flows, fixed costs and public funds together, in two regimes. In the short run the number of frontier providers is fixed: entry is blocked by compute, talent and capital, providers earn rents, and fixed costs are sunk. In the long run entry is free and profit is competed away. The two regimes disagree about ownership and, under constant-elasticity demand, agree about the sign.

First: the sign is not a claim about AI. A markup does two things and they take two instruments. It transfers surplus to shareholders, which is inframarginal and is reached by a tax on profits. It also wedges price above marginal cost, so the good is consumed too little, and that is allocative and is reached only on the margin. A usage tax is an instrument of the second kind applied to a problem of the first kind: it does not reach the rent, which is inframarginal, and it deepens the quantity distortion the markup already created. So a good sold above marginal cost is consumed too little, and the textbook corrective for underconsumption is a lower price. Equation (8) inherits that from Cournot (1838) and would return the same sign for bread. The formula prices underprovision on the only margin it has been given, because Proposition 2 solves for a single wedge. It does not know that the planner also has an antitrust authority and

a procurement policy, which are the instruments for the markup itself; a cash flow tax collects the rent but, being neutral, leaves the markup and the underprovision in place. Once competition policy does its job the term leaves the usage tax's job description entirely. Nothing in this paper recommends paying frontier labs; Appendix C.1 says so at length. What the term does is measure the size of one distortion relative to another, and that comparison is informative whatever one thinks of the instrument.

Second: with entry blocked, the term is weighted profit. Proposition 2 weighted provider profit at one. The margin on a marginal unit is profit, so its proper weight is the planner's weight on the holders of that profit.

Proposition 7 (Fixed providers)**.** *With the number of providers fixed and constant-elasticity demand, and with provider profit weighted by g_S relative to public funds,*

$$t_U^\star \;=\; \frac{\Omega}{p} + \frac{D-\psi}{p} \;-\; (1-f)\, g_S\Big(1-\frac{1}{\mu}\Big).$$

The shareholder weight is not a free parameter. A planner who values a dollar to routine workers above the average must value a dollar to the holders of frontier-lab equity below it, and the two are tied by the same welfare function.

Lemma 4 (The shareholder weight implied by $\hat{g}_1$)**.** *Let welfare weights be isoelastic in income, $g(y) \propto y^{-\eta}$, normalized so that the income-weighted mean weight is one, and let log income be normal with dispersion s. Then $g(y) = \exp\big(-\eta s z + \frac{1}{2}\eta(2-\eta)s^2\big)$, where z is the income z-score. The requirement $g = 1.3$ at the routine worker's position has two roots in η, about 0.32 and 2.55; we take the smaller, since the larger implies an inequality aversion under which shareholders receive almost no weight ($g_S \approx 0.006$), and it pins η; evaluating g at the equity-weighted holder's position then delivers g_S.*

Two observable positions discipline the calibration. Routine occupations sit near the 36th percentile of the earnings distribution and corporate equity is held overwhelmingly at the top, which together give $g_S \approx 0.66$, with a range of 0.45 to 0.85. Appendix B.9 gives the derivation, shows that, at the income positions we assume, $\hat{g}_1 = 0.3$ and $g_S = 0.2$ are not jointly available under an isoelastic welfare function, and answers the objection that the deadweight loss from market power might simply be smaller than that from a tax.

The discipline matters because the conclusion is sensitive to it. Kleven and Zidar (2026) set the welfare weight on the owners of capital to zero, which in their model makes

the optimal capital tax the revenue-maximizing rate $1/(1+\zeta_K)$. A planner who weights routine workers only thirty percent above the average cannot consistently do so here: $g_S \approx 0.66$ is what that consistency requires, and a lower g_S would shrink the market-power term with entry blocked and so raise the optimal usage wedge. Their capital tax is also, formally, a uniform tax on machine inputs, and in their competitive economy it is equivalent to a tax on the income of machine owners. With a marked-up supplier the equivalence fails. A tax on usage then also deepens the underprovision this section describes, whereas a tax on rent does not, which is why the two instruments have to be kept apart here.

Third: under free entry, ownership drops out. When entry is free, the margin over marginal cost is not rent: it pays for the fixed cost of the marginal entrant, and profit is zero. A usage tax still has a first-order cost beyond the revenue it raises, because it raises the markup. Buyers lose Φ per unit of tax on every unit they buy, while the government collects one.

Proposition 8 (Free entry)**.** *Under symmetric Cournot competition with free entry and constant-elasticity demand, provider profit is identically zero, so the market-power term carries no shareholder weight and does not depend on ownership. To first order,*

$$t_U^\star \;\approx\; \frac{\Omega}{p} + \frac{D-\psi}{p} \;-\; \frac{\Phi-1}{\sigma_q\,\Phi},$$

with Φ as in Proposition 6, and the rate is the same for a domestic and an importing economy.

The gain from lower prices accrues to buyers and taxpayers, which is why ownership does not enter: Corollary 2 holds only while entry is blocked, and in the long run an importer gains from cheaper AI exactly as a domestic economy does. The first-order expression overstates the term at high markups, because Φ rises as the tax shrinks the market, so the calibration solves the equilibrium exactly. Measured against the textbook term $1-1/\mu$, free entry absorbs roughly half of the correction for an ad valorem tax, a share that rises from 0.47 at $\mu = 1.2$ to 0.56 at $\mu = 2$ (Appendix B.6).

Fourth: which regime, and what recurring training tells us. Frontier training recurs: each model generation requires a new run, and the run is the price of staying at the frontier. That fact does not shrink the market-power term. It tells us which regime applies at which horizon. Within a model generation, the number of frontier providers

is effectively fixed and the fixed-provider regime describes the response to a tax. Across generations, recurring fixed costs are what free entry competes over, and the free-entry regime is the relevant benchmark. Reporting both is therefore not hedging between rival models but distinguishing two horizons of the same industry. Neither regime supports a positive usage tax for a domestically owned sector at any markup above about ten percent.

5.7 Diffusion spillovers, and the general formula

If a marginal unit of AI use generates learning or diffusion benefits the buyer does not capture, the Pigouvian term is netted of them. The sign of these spillovers is unsettled in the literature, so we carry them as a parameter ψ of either sign; positive spillovers lower the optimal tax and negative ones raise it. Appendix B.12 develops the point. Collecting the results of this section, the general formula is

$$t_U^\star \;=\; \underbrace{\frac{(1-\tau)\hat{g}_1\Delta_\ell(\sigma-1)\beta_1}{(1+\varepsilon_L+\mathcal{E})\sigma_q}}_{\text{predistribution}} \;+\; \underbrace{\frac{D-\psi}{p}}_{\text{net externality}} \;-\; \underbrace{M}_{\text{market power}}, \qquad M = \begin{cases} (1-f)\,g_S\,(1-1/\mu) & \text{entry blocke} \\ [2pt](\Phi-1)/(\sigma_q\Phi) & \text{free entry, to} \end{cases} \tag{9}$$

with the training margin handled by a separate instrument $t_K^\star = R'(K)$. Proposition 2 is the blocked-entry case with $g_S = 1$ and $\psi = 0$; fixed assignment is a further special case. The extensions do not all move the formula the same way. The earnings base raises the predistribution term relative to the task-income version, the exposure gap and reallocation lower it, a shareholder weight below one shrinks the market-power term, and free entry replaces it with an amplification term that does not depend on ownership. Under free entry the markup responds to the tax, and the calibration solves the equilibrium exactly rather than using the first-order expression.

6 Calibration

The formula is only as useful as the numbers in it. This section calibrates each term, then runs five back of the envelope calculations through the model's own lemmas, then reads the policy debate against the results. Parameters are chosen, where there is doubt, to favor the case for taxing AI.

6.1 The three terms

The residual external cost per kilowatt-hour. Section 2 gave the two components. Unpriced carbon at \$190 per ton and $\zeta = 0.4$ kg per kWh is \$0.076 per kWh where no carbon price applies; the grid-shifting wedge γ we carry over $[0, 0.03]$; water and local damages add a few tenths of a cent. A central residual, with a carbon price of roughly \$100 per ton already in place and γ near the middle of its range, is

$$d \approx \$0.05 \text{ per kWh}, \qquad \text{range } \$0.02 \text{ to } \$0.10,$$

and higher with less carbon pricing: if the effective price on U.S. data-center load is below \$100 per ton, both d and the Pigouvian term are larger, though the case for levying the charge on electricity is unchanged. At the central d the residual wedge is therefore about sixty percent of the price of the input on which it should be levied, and above one hundred percent at the top of the range. By the standards of environmental taxation that is an extraordinary number: it says the uncorrected externality is of the same order as the price of the electricity itself. Figure D.2 shows it is robust across the whole range of γ, because carbon dominates. This separates the two roles of γ: it barely moves the *level* of the wedge, and what it governs instead is the distributional character of the status quo, which is Section 6.6.

The residual external cost per unit of AI, across workloads. Here the point estimate is misleading and the distribution is the object that matters. A short text query consumes on the order of 0.3 Wh, while reasoning and agentic workloads run one to two orders of magnitude higher (International Energy Agency, 2026), and these are where load growth is concentrated. The broader pattern is long established: a generative query uses roughly ten times the energy of a conventional web search (de Vries, 2023), and multi-purpose generative models use orders of magnitude more energy per task than task-specific ones (Luccioni et al., 2024). Because both energy and price scale with query length, D/p is roughly constant *within* a workload type but differs *across* types through the price per joule of delivered value. Table 1 works the ratio for four representative workloads at $d = \$0.05$ per kWh.

The two means differ by a factor of three, and the energy-weighted mean is the policy-relevant one, because a corrective instrument should track the harm, which is proportional to energy. For a uniform ad valorem tax the relevant benchmark is total damage over total expenditure, which Table 1 puts at 3.1%; the count-weighted 1.3% understates it

Table 1: The externality ratio D/p varies across workloads

Workload	Energy/query	Price/query	D/p	Share of queries	Share of energy
Simple text	0.3 Wh	\$0.003	0.5%	0.60	0.02
Standard chat	3 Wh	\$0.010	1.5%	0.18	0.05
Reasoning	30 Wh	\$0.050	3.0%	0.19	0.52
Agentic	150 Wh	\$0.150	5.0%	0.03	0.41
Count-weighted mean			1.3%		
Energy-weighted mean			**3.7**%		

Notes. Energy and price per query are order-of-magnitude figures drawn from International Energy Agency (2026), Luccioni et al. (2024) and de Vries (2023); D/p evaluates $d \cdot e/p$ at $d = \$0.05$ per kWh. The query mix is illustrative rather than measured: no public source reports the composition of inference by workload type, so we impose a plausible mix in which short queries dominate by count while reasoning and agentic workloads dominate by energy. The share of energy is derived from the query shares and energy per query, not assumed separately. The two means are the D/p column weighted by the two share columns respectively. Nothing in the paper's argument depends on the particular mix, only on the fact that the ratio of harm to value rises across the rows.

by more than half, because the energy-hungry tail dominates the footprint. The token tax therefore not only sits on the wrong base, but on one whose relationship to the harm is itself unstable across the workloads driving the load growth. We carry the energy-weighted central value $D/p \approx 2$–4% forward, and note it would be higher still with no carbon price.

Market power. Gross margins on frontier inference of fifty to sixty percent are often read as a markup of 2 to 2.5.[15] The model limits how far that reading can go. No firm can face less elastic demand than the market it shares, so the Lerner index satisfies $1 - 1/\mu \leq 1/\sigma_q$ under any symmetric conduct, homogeneous or differentiated, Cournot or Bertrand; at $\sigma_q = 1.965$ this caps the markup at 2.04. A fifty percent margin is therefore consistent with the model only under near-monopoly conduct, and a sixty percent margin not at all: either frontier provision behaves as a near-monopoly, or reported gross margins overstate the markup over marginal cost. We parametrize conduct by $h \in (0, 1]$, with $1 - 1/\mu = h/\sigma_q$ (the Cournot Herfindahl), take $\mu = 1.5$ as central, between duopoly and monopoly conduct, and report a structure-based $\mu = 1.2$, about three providers, and a margin-consistent $\mu = 2.0$. The shareholder weight is $g_S = 0.66$ from Lemma 4, so with entry blocked the market-power term is $g_S(1 - 1/\mu) \approx 22$ percent at the central markup.

[15] These margins come from provider disclosures and investor communications rather than from audited segment accounts, and inference is not reported as a separate segment by any major provider. Allocating shared training, capacity and research costs between inference and other activities involves judgment we cannot verify from outside.

Predistribution. The formula has three inputs and only one of them is normative. Take $\tau = 0.35$, $\varepsilon_L = 0.3$, $\sigma = 2$, and a welfare weight $\hat{g}_1 = 0.3$, so that the planner values a dollar to the exposed group thirty percent above the average. At the fixed-assignment corner $\Delta_\ell = 1$, $\mathcal{E} = 0$ (the configuration the robot-tax literature works in) the exact term of (8), with $(1-s)$ in the numerator and $\sigma_q = \sigma(1-s)+s+\theta \approx 1.965$ in the denominator, gives $\Omega/p \approx 7.3\%$. We report that number only as what it is, the value the term takes at $\Delta_\ell = 1$, which is the corner of the parameter space the robot-tax literature imposes by assumption and the corner at which the term is largest.

The exposure gap is not one. Δ_ℓ is a difference of two employment shares—the share of the disadvantaged ex ante group working in AI-substitutable occupations, less the share of the advantaged group—so it requires an exposure measure, a grouping, and a threshold, and given those three it is a descriptive statistic computable from one year of the American Community Survey. It is not a correlation, not a regression coefficient, and not bounded below by zero.[16]

We compute it. Table 2 reports Δ_ℓ for seven public occupation-level exposure measures crossed with four ex ante groupings, on an employment-weighted panel of 765 detailed SOC occupations; Appendix D.9 gives the construction and its limitations. Three findings come out of the table.

First, the exposure gap is nowhere close to one. The largest absolute value in the table is 0.50 and the largest positive value is 0.45; the left panel of Figure 1 plots these against the value the literature assumes. The predistribution term is proportional to Δ_ℓ and is damped further by reallocation, so the convention $\Delta_\ell = 1$ overstates it by about a factor of two at the most favourable measure in the table, 7.3 percent against 3.6, and gets its sign wrong at every direct AI measure.[17] The factor is two rather than larger because the corner is a different economy as well as a different gap: in the two-type model group 1 is the routine workforce, whereas in the calibrated economy the group the planner weights earns more than routine income, which partly offsets the smaller gap.

Second, the measures disagree about the sign, and they disagree by family. The robot-exposure measure of Webb (2020) is the closest available proxy for the routine displacement that σ describes, and it gives $\Delta_\ell = +0.41$ for a less-than-BA group 1, with $+0.34$ to $+0.45$ across the other groupings. The software measure of the same paper gives a uniformly smaller $+0.18$ to $+0.23$. On all four direct AI and generative-AI measures,

[16] Appendix D.9 states the estimand formally, and the replication package includes the code.

[17] The task-based simulations of Kleven and Zidar (2026) point the same way from the other side: calibrated to the U.S. wage distribution, automation intensity peaks near the 41st wage percentile rather than at the bottom, so the workers most exposed are not the lowest paid.

Felten et al. (2021), Eloundou et al. (2024) under both the human and the GPT-4 rubric, and Eisfeldt et al. (2023), the gap is negative for every grouping, between -0.18 and -0.50. The split has a simple reading. Robots and routine software substitute for codifiable tasks concentrated in middle- and lower-skill occupations, so their exposure falls on the less educated and the gap is positive. Generative AI reaches non-codifiable cognitive work (Ide and Talamàs, 2025), and on the direct AI measures that work is concentrated in high-education occupations, so measured exposure falls on the more educated and the gap turns negative. This is why negative gaps are rare in the robot-tax literature: the technology it studies is of the first kind. The employment-weighted correlation between exposure and the share of an occupation's incumbents without a bachelor's degree is $+0.52$ on the robot measure and -0.25 to -0.64 on the four AI measures. The choice of grouping does not move the sign. The two families of measure describe different technologies.

Evidence of a different kind supports the negative side of the split. Tomlinson et al. (2025) do not rate occupations for exposure. They classify two hundred thousand logged conversations with a deployed assistant into the work activities it was asked to perform, then aggregate to occupations. Their applicability score correlates with occupational wages at $+0.13$ when occupations are weighted by employment, and it spreads broadly across education requirements. This is what our negative Δ_ℓ also says, obtained without any expert judgment about what a model could do. We weight it more heavily than we would agreement among rating-based measures, which share a method and could share its errors.

Third, the gap shrinks as the definition of exposure is tightened. Figure D.3 sweeps the threshold from the fortieth to the ninetieth percentile of employment. Six of the seven measures collapse toward zero, and at the ninetieth percentile all seven lie between -0.22 and $+0.10$.[18] This fall is partly mechanical, however, and should not be read as evidence about segregation. With exposed employment share E and group-1 share π, the gap cannot exceed $\min\{E/\pi, (1-E)/(1-\pi)\}$, so tightening the cut drives it toward zero even under perfect segregation. The informative numbers are the gaps at the stipulated cuts in Table 2.

We carry $\Delta_\ell = 0.4$ as the central case. This is the value on the single measure most favourable to the robot-tax case, and we carry it for that reason, following this section's practice of resolving doubt in favour of taxing AI. The range the measurement supports is $[-0.50, +0.45]$, which contains zero.

Remark 2 (The sign of Δ_ℓ is not guaranteed). Every treatment of the robot tax we know

[18] The exception is Webb software, which stays flat inside a band around $+0.1$ at every cut.

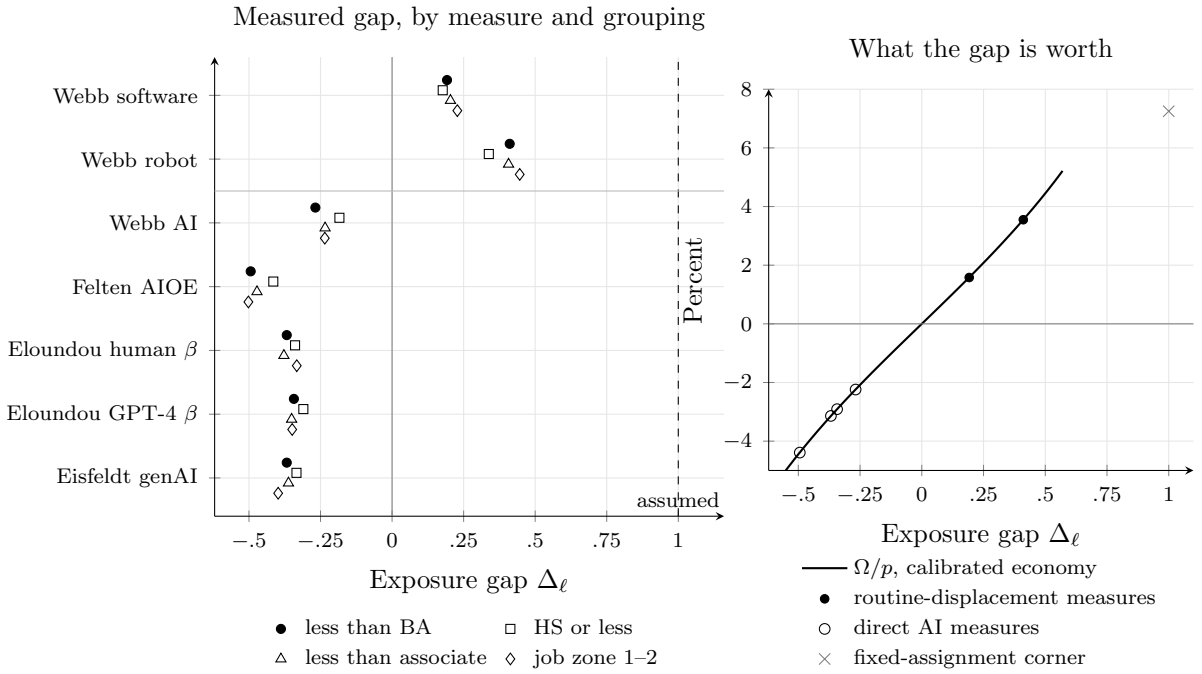


Figure 1: The exposure gap, measured, and what it is worth. Left: Δ_ℓ for seven public occupation-level exposure measures crossed with four ex ante groupings, from Table 2; the horizontal rule separates the two routine-displacement measures above from the four direct AI measures below, and the dashed vertical line is the value $\Delta_\ell = 1$ that the robot-tax literature imposes by assumption. One of the twenty-eight cells reaches half that value in absolute terms and none is positive beyond 0.45. The four markers move together within every row, so the sign is a property of the measure and not of the grouping. Right: the predistribution term Ω/p of (8) as a function of Δ_ℓ in the calibrated economy, with equal-sized groups, routine work at 29 percent of labor income, group 1 earning 45 percent of it, and $\mathcal{E}$ from Lemma 2 at $\kappa = 1.5$; other parameters as in Table 4. With equal groups and that routine share, gaps beyond 0.58 are infeasible, so the curve stops there. The fixed-assignment corner, marked separately, is the two-type case in which group 1 is the routine workforce. Markers give the measured values at the less-than-BA grouping.

of imposes $\Delta_\ell = 1$ and therefore cannot express the possibility that the exposure gap is negative. Nothing in the theory rules it out, and Table 2 finds it to be the case. On all four direct AI and generative-AI measures, and on all four groupings, Δ_ℓ lies between -0.18 and -0.50. When $\Delta_\ell < 0$ the predistribution term in (8) becomes negative, so both it and the market-power term push the optimal wedge down. They do so for unrelated reasons, and it is worth keeping them apart: the market-power term is a positive magnitude entering with a minus sign, because a good priced above marginal cost is consumed too little; the predistribution term keeps its plus sign and turns negative in value, because the group a planner would weight more heavily is the one *less* exposed to the technology. A usage tax then *worsens* the distribution across ex ante groups, because it protects the tasks performed disproportionately by the better-off, and the redistributive argument turns into an argument against the instrument. At the measured LLM gap of -0.37 the term is -3.1 percent against the $+7.3$ percent of the fixed-assignment corner, a swing of more than ten points on the same formula. Which sign applies depends on what

Table 2: The exposure gap Δ_ℓ, measured

	Ex ante grouping (group 1 = less educated)				
Exposure measure	$<$ BA	HS or less	$<$ assoc.	job zone 1–2	Ω/p
Webb (2020) software	+0.19	+0.18	+0.20	+0.23	+1.6
Webb (2020) robot	**+0.41**	+0.34	+0.41	+0.45	+3.6
Webb (2020) AI	−0.27	−0.18	−0.23	−0.24	−2.2
Felten et al. (2021) AIOE	−0.49	−0.42	−0.47	−0.50	−4.4
Eloundou et al. (2024), human β	−0.37	−0.34	−0.38	−0.33	−3.1
Eloundou et al. (2024), GPT-4 β	−0.34	−0.31	−0.35	−0.35	−2.9
Eisfeldt et al. (2023) genAI	−0.37	−0.33	−0.36	−0.40	−3.1
Memo: the literature's convention	+1.00	+1.00	+1.00	+1.00	+7.3

Notes. $\Delta_\ell = \ell_{1R} - \ell_{2R}$, the employment-weighted share of group 1 in exposed occupations less that of group 2. Group 1 is the less-educated or lower-preparation group in each column. An occupation is exposed if its score is above the employment-weighted 66th percentile for the two Webb (2020) routine-displacement measures, the Autor–Dorn convention, and above the employment-weighted median otherwise. The panel is 765 detailed SOC-2018 occupations; within-occupation education distributions are the incumbent-reported Required Level of Education from O*NET 30.3 (National Center for O*NET Development, 2026), and employment weights are Dorn (2009) labor-supply weights allocated to SOC. The last column is the predistribution term Ω/p of (8) in percent, at the $<$ BA gap, with $\tau = 0.35$, $\varepsilon_L = 0.3$, $\sigma = 2$, $\hat{g}_1 = 0.3$, $\kappa = 1.5$, equal-sized groups and the calibrated earnings base, so that $\mathcal{E}$ moves with each row. Appendix D.9 reports the threshold sweep and the construction in full.

σ describes. If it is routine displacement then the robot measure is the right one and $\Delta_\ell \approx +0.4$; if it is direct LLM substitution then it is not. We do not settle that here, and the frontier is moving in any case. What matters is that the question is empirical and has a measured answer under either reading, and that a literature imposing $\Delta_\ell = 1$ cannot pose it.

Given Δ_ℓ, the reallocation elasticity is not a free parameter: Lemma 2 pins it in efficiency units. At the Roy–Fréchet estimates of Hsieh et al. (2019) and Galle et al. (2023), $\kappa \approx 1.5$ to 2; with equal-sized groups, routine work at $h = \lambda(1-s)/(1-\lambda s) = 0.29$ of labor income and $\Delta_\ell = 0.4$, this gives $\mathcal{E} \approx 0.40$ to 0.81. The earnings base matters as much. In the symmetric benchmark of Proposition 5 group earnings and routine earnings coincide, but in the calibrated economy routine work is only 29 percent of labor income, so a group earning 45 percent of it has an earnings base $b \equiv B_1/(w_R L_R) \approx 1.56$ times routine income. Combining,

$$\Omega/p \;\approx\; 3.4\%, \qquad \text{range } 2.5\% \text{ to } 3.8\%,$$

over $\kappa \in [1.5, 2]$ and a group earnings share between 40 and 50 percent. This is about half the 7.3 percent of the fixed-assignment corner: the exposure gap cuts the corner value by more than half and reallocation trims it further, while the broader earnings base restores part of the loss. It sits within the one to four percent that Costinot and Werning (2023)

Table 3: The optimal usage tax, by markup and regime

			Entry blocked		Free entry
Markup μ	firms n	Φ	domestic	importer	any ownership
1.2	3.05	1.22	−4.5%	+6.4%	−2.4%
1.5	1.53	1.65	−15.5%	+6.4%	−10.2%
2.0	1.02	2.90	−26.5%	+6.4%	−15.6%

Notes. Central calibration, $(\Omega + D)/p = 6.44\%$ and $g_S = 0.66$. The number of firms is the Cournot count implied by μ at $\sigma_q = 1.965$, and Φ the free-entry pass-through of Proposition 6. Entry blocked: Proposition 7, exact. Free entry: the optimum over the solved equilibrium (Appendix B.6). The domestic optimum is negative in both regimes at any markup above 1.11 (entry blocked) and 1.14 (free entry).

Table 4: The optimal AI usage tax, by scenario (percent)

	Predistr.	Extern.	Mkt. power	Wedge $t_U^\star$	
Scenario	(1)	(2)	(3)	(1)+(2)	(1)+(2)+(3)
Baseline, domestic	+3.4	+3.0	−21.9	+6.4	**−15.5**
Baseline, importer	+3.4	+3.0	+0.0	+6.4	**+6.4**
Free entry, any ownership	+3.4	+3.0	−16.6	+6.4	**−10.2**
Structure-based markup (μ=1.2)	+3.4	+3.0	−11.0	+6.4	**−4.5**
Margin-consistent markup (μ=2.0)	+3.4	+3.0	−32.9	+6.4	**−26.5**
Emissions priced at source	+3.4	+0.0	−21.9	+3.4	**−18.5**
No shareholder discount (g_S=1)	+3.4	+3.0	−33.3	+6.4	**−26.9**
Measured robot gap (Δ_ℓ=+.41)	+3.6	+3.0	−21.9	+6.6	**−15.4**
Measured LLM gap (Δ_ℓ=−.37)	−3.1	+3.0	−21.9	−0.1	**−22.1**
Fixed-assignment corner	+7.3	+3.0	−21.9	+10.3	**−11.7**

Notes. Central calibration: $\mu = 1.5$, $\sigma = 2$ ($\sigma_q = 1.965$), $\hat{g}_1 = 0.3$, $g_S = 0.66$, $\Delta_\ell = 0.4$, $\kappa = 1.5$, equal-sized groups with group 1 earning 45 percent of labor income, $D/p = 3\%$, $\psi = 0$. Rows are for a domestically owned sector with entry blocked unless stated. The free-entry row reports the exact optimum, with column (3) the implied market-power term. The measured-gap rows evaluate column (1) at the values of Table 2. The fixed-assignment corner is the two-type case used throughout the robot-tax literature, with group 1 the routine workforce.

report for industrial robots.

Table 4 collects the results, and its most striking feature needs addressing head-on rather than left for the reader to notice.

6.2 Why the market-power term dominates

Table 4 puts every domestic entry of the last column below zero, and Table 3 shows what governs them. Both deserve a plain reading, because a column of negative numbers invites the objection that the model recommends subsidizing the most profitable firms in the economy, and it does not.

The arithmetic is simple. Predistribution and Pigouvian terms sum to about 6.4 percent at the baseline and are positive in every row with a positive exposure gap. The

market-power term is $g_S(1 - 1/\mu) \approx 22$ percent with entry blocked and about 17 percent under free entry. It dominates, at any markup above about ten percent. The substantive point is that *a single wedge on the usage margin is being asked to do two incompatible jobs*: to correct an externality and a wage distortion, both of which want a higher price of AI, and to correct underprovision, which wants a lower one. At defensible parameters the second is the larger.

The right response is not to accept the residual as a recommendation but to stop asking one instrument to do both. A cash flow tax collects the rent but, being neutral, leaves the markup and the underprovision in place. The instruments for the markup itself are entry, interoperability and procurement policy, which we do not model. If they do their job, column (3) is no longer the usage tax's, and column (1)+(2), a positive wedge of about 6.4 percent at the baseline, becomes the operative number for a finance ministry. On the same draws as Section 6.4, $\Pr\big(\Omega/p + (D - \psi)/p > 0\big) = 1.00$, with a median of 5.4 percent, and that wedge exceeds ten percent in 5 percent of draws. The two readings differ in sign but agree that the rates under discussion sit at or above the top of what the model supports.

6.3 Three calculations

Table 4 prices the tax. The model also prices the things the tax is supposed to achieve, and the orders of magnitude are not what the debate assumes. Throughout we take exposure $\theta = \lambda s = 1.5\%$ of GDP, which is generous given that global spending on AI services is currently well under one percent of world output, and we set $\lambda = 0.3$ and $s = 0.05$.

1. What a usage tax raises, and what it costs. A ten percent tax on a base worth 1.5 percent of GDP raises 0.15 percent of GDP before behavioral response and 0.124 percent after it. With full welfare accounting, which counts the pre-existing markup, the externality and the wage effects as well as the tax, it lowers domestic welfare by 0.052 percent of GDP with entry blocked and 0.087 percent under free entry: 0.42 and 0.75 dollars of welfare per dollar raised. A fifty percent cash flow tax on the rents of providers whose entry is blocked raises about 0.25 percent of GDP at no first-order cost, but under free entry there is no rent for it to reach. And at $d = \$0.05$ per kWh the emissions charge has a base of \$9 billion today on U.S. data center load of roughly 180 TWh, rising to \$20–30 billion by 2030 (Shehabi et al., 2024; International Energy Agency, 2026). Section 7 scores all four instruments.

2. What a usage tax does for routine workers. By Lemma 2, a ten percent usage tax raises the routine wage relative to the non-routine wage by

$$\frac{s(\sigma-1)}{1+\varepsilon_L+\mathcal{E}} \times 10\% \;=\; \frac{0.05 \times 1}{1.70} \times 10\% \;\approx\; 0.3\%.$$

Three tenths of one percent, and that is the movement in the relative *task* wage at $\mathcal{E} = 0.40$; at the fixed-assignment corner, where $\mathcal{E} = 0$, it would be 0.38 percent. What the exposed *group* gains is smaller again by the exposure gap: $\Delta_\ell \times 0.29 \approx 0.12$ percent. Keeping these two objects apart matters, because the first is what an incidence regression would recover and the second is what enters welfare. This is the entire predistributive payoff of a policy whose net welfare cost Calculation 1 puts at 0.42 dollars per dollar raised, and it accrues as a relative wage movement rather than as income to any identifiable person. An earned income tax credit of similar fiscal size would plausibly deliver more to the same workers, without touching the price of compute or depending on σ. Our model does not include such a credit, but Kleven and Zidar (2026) derive the comparison in a task-based model with complete labor taxes and find that the optimal response to automation is a substantially more generous earnings subsidy at the bottom, not a tax on machines.

There is reason to think even this is an upper bound. Saez et al. (2019) study a large, long-lasting employer payroll tax cut for young workers in Sweden and find no effect on their net-of-tax wages even six years on.[19] The mapping is not exact, since their variation taxes labor rather than a substitute for it, but the direction is uncomfortable for the predistribution term. A model in which a factor wedge's incidence is fully expressed in relative wages is the model that maximizes the term in $\hat{g}_1$; if much of it instead shows up in quantities or in firm rents, the wage compression a usage tax buys is smaller still. We flag this as a bound rather than a correction. It holds within a model of voluntary labor supply and costless sorting; with employment constraints, costly switching or lost task-specific human capital, quantity responses carry welfare costs of their own and the direction is no longer guaranteed.

3. Why the wedge is on the wrong base. Take the standard-chat workload, energy 3 Wh and price one cent. Electricity is $p_e e/p = 2.4\%$ of the price of the query, and the residual external cost is sixty to a hundred percent of the price of that electricity, so the external cost of the query is between 1.4 and 2.4 percent of its price. Read backwards the same fact is more striking. A usage tax at a politically meaningful rate, say ten

[19] The response appears as employment, and the windfall accrues to firms that happen to employ many young workers, a rent-sharing channel and not a relative-price one.

percent, applies a wedge of \$0.001 per query against an external cost near \$0.00015: it over-corrects several-fold for a typical chat query, and by Table 1 it over-corrects for every workload, from twofold for agentic tasks to twentyfold for simple text. A charge of \$0.05 per kWh at the data center prices the harm exactly for every workload at once and would be invisible on any invoice. The point is not merely that the token tax mis-sets the level; it is that no single ad valorem rate can track a harm whose ratio to value varies by an order of magnitude across workloads, whereas a per-kilowatt-hour charge tracks it by construction. That is the whole argument in one line, and the reason is arithmetic rather than ideological: an ad valorem tax on an information good is a poor instrument for a physical externality.

Two further calculations are relegated to Appendix D: what a ninety percent fall in the price of AI does to the wage distribution, which bears on the size of the problem rather than on the instrument; and the welfare cost of choosing the wrong instrument to raise a given sum, which is about eight cents of deadweight per dollar for the usage tax against zero to first order for the rent tax and a first-order gain for the emissions charge.

Appendix B.7 compares the four instruments on revenue, distortion and output.

6.4 A probabilistic calibration

Table 4 reports scenarios, and a scenario table invites the objection that the conclusion is an artifact of one column. Since several parameters are uncertain at once, we place uniform priors on all of them and report the induced distribution of the optimal tax.[20] We draw $\sigma \sim U[1.2, 3]$, conduct $h \sim U[0.3, 1]$, $\hat{g}_1 \sim U[0.1, 0.5]$, $\varepsilon_L \sim U[0.1, 0.5]$, $\kappa \sim U(1, 3]$, $\Delta_\ell \sim U[0.2, 0.6]$, group 1's employment share $\pi \sim U[0.5, 0.62]$ and earnings share $\sim U[0.40, 0.50]$, $d \sim U[0.02, 0.10]$, an energy share of price $\sim U[0.02, 0.06]$, and $\psi/p \sim U[-0.015, 0.015]$, with 200,000 draws; Appendix D.4 reports the full distribution and a correlated-prior variant.

Four restrictions are imposed draw by draw rather than left to independent priors. The markup satisfies the Lerner bound by construction, $1 - 1/\mu = h/\sigma_q$, so every draw is a feasible equilibrium. The reallocation elasticity is computed from Lemma 2, so no draw combines a mobility elasticity and an exposure gap that are jointly infeasible, and exposure gaps larger than the group and routine shares allow are discarded (20 percent of raw draws). The shareholder weight is derived from $\hat{g}_1$ through Lemma 4, drawing the income positions behind it, so the two weights are never combined inconsistently. And $\kappa > 1$ throughout, so the expected wage is finite. The free-entry optimum is interpolated

[20] This is not an estimate; it is a statement about how much of the parameter space the conclusion occupies.

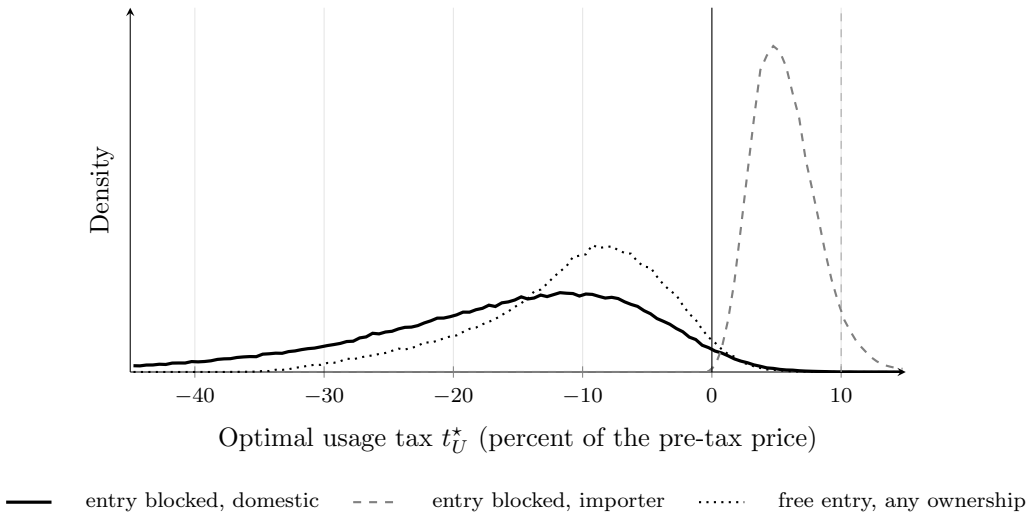


Figure 2: The distribution of the optimal usage tax under the priors of Section 6.4, 200,000 draws. No distributional form is imposed: every prior is uniform and each curve is a histogram of the simulated draws. With entry blocked, the domestic optimum has median −14.8 percent and is negative in 97 percent of draws; the importer's has median +5.4 percent and exceeds ten percent, the dashed line, in 5 percent. Under free entry ownership drops out, and the optimum has median −9.7 percent, negative in 97 percent of draws. All three rely on constant-elasticity demand; Section 6.5 shows how they change when providers absorb part of a tax.

from a grid of exact equilibrium solutions, accurate to 0.14 percentage points.

Figure 2 plots the result. For a domestically owned sector the optimal usage tax is negative in 97 percent of draws with entry blocked and in 97 percent under free entry, and above five percent in under one percent of draws in either regime. For an importer with entry blocked it is positive in essentially every draw, with median 5.4 percent; under free entry the importer's optimum coincides with the domestic one. The predistribution term has median 2.2 percent and stays below five percent in 90 percent of draws.

One prior deserves comment. The draw $\Delta_\ell \sim U[0.2, 0.6]$ is one-sided in a way Table 2 does not support, and we keep it because it is the configuration favourable to the instrument. Widening it to the measured $U[-0.50, 0.45]$, and leaving every other prior alone, moves the predistribution term to a median of -0.1 percent, with a fifth percentile of -4.0 and a ninety-fifth of $+3.5$, and takes $\Pr(\Omega/p < 0)$ from zero to 0.52. Every probability we report moves against the usage tax under the measured range.

Appendix D.4 reports the full distribution of each term and a robustness check in which the parameters are drawn from a Gaussian copula rather than independently, with $\rho(\sigma, \hat{g}_1) = 0.3$, $\rho(d, \text{energy share}) = 0.3$ and $\rho(\kappa, \Delta_\ell) = -0.4$, the last because sharp sorting and fast reallocation are substitutes even beyond the deterministic frontier. Correlation fattens the lower tail but leaves the medians close to unchanged, and moves the domestic probability of a negative optimum only from 0.97 to 0.96 with entry blocked.

6.5 Pass-through and the sign of the tax

Every result so far uses constant-elasticity demand, which is what the task structure of Section 3 generates. Under it, the producer price does not respond to a usage tax when entry is blocked, and a cost increase is passed through at μ. Both properties are specific to the demand system, and the sign of the optimal usage tax turns on them. The mechanism can be stated exactly. Let $\rho \equiv \mathrm{d}\ln q / \mathrm{d}\ln(1+t_U)$ be the pass-through of the usage tax into the buyer price. Generalizing Proposition 7 to any demand curve (Appendix D.5), the optimal tax satisfies, to first order,

$$t_U^\star \;\approx\; \frac{\Omega + D - \psi}{p} \;-\; (1-f)\, g_S\Big(1 - \frac{1}{\mu}\Big) \;+\; \big[1 - (1-f)\, g_S\big]\, \frac{1-\rho}{\sigma_q\, \rho}. \tag{10}$$

The first two terms are those of Proposition 7. The third is the value of rent extraction. When providers absorb part of the tax, $\rho < 1$, the tax converts provider profit into public revenue, and that conversion is worth the gap between the value of public funds and the planner's weight on the profit, $1-(1-f)g_S$: about a third for a domestic owner, and one for an importer, whose providers' profit counts for nothing domestically. Under constant-elasticity demand $\rho = 1$ and the term vanishes, which is why the earlier results take the form they do.

The same expression covers free entry. There profit is zero, so the underprovision term has nothing to weight, and $\rho = \Phi > 1$: the tax drives out entrants and raises the buyer price by more than itself. The third term then turns negative, $(1-\Phi)/(\sigma_q \Phi)$, which is Proposition 8. In both regimes the market-power logic runs through pass-through: below one, a tax extracts rent; above one, it creates markup.

Linear demand is the natural contrast. Calibrated to the same markup and the same elasticity at the initial equilibrium, linear Cournot demand with entry blocked passes through about 0.60 of a cost increase in levels at $\mu = 1.5$, which in the proportional terms relevant to an ad valorem tax is $\rho \approx 0.40$: buyers bear forty percent of the tax and providers sixty. The rent-extraction term is then about +26 points for a domestic owner and +75 for an importer, against an underprovision term of −22. Rent extraction wins, and the optimal usage tax becomes *positive*: about 8 percent for a domestically owned sector and 40 percent for an importer at the exact optimum, compared with −15.5 and +6.4 percent under constant elasticity. The first-order expression gets the sign right but overstates the importer's rate, because pass-through changes as the tax grows. At $\mu = 1.2$ the linear-demand optima are about 6 and 23 percent, and at $\mu = 2$ about 17 and 64 percent. The emissions charge also reverses: under linear demand it raises domestic

welfare slightly and an importer's by more, because providers absorb part of it.

What value of ρ is plausible? Equation (10) turns the question of sign into a question about one number. At the central calibration the domestic optimum stays negative for any ρ above about 0.53; below that, rent extraction outweighs underprovision. The importer's optimum is positive for any $\rho \leq 1$ and rises steeply as ρ falls. Table 5 reports both at the values the evidence suggests.

Measuring ρ directly at the frontier is not possible with public data. Lemma D.1 shows that ρ equals the elasticity of the buyer price with respect to an industry-wide shock to marginal cost, so in principle it can be read off how prices respond to accelerator rental rates. For closed models the cost side cannot be measured: marginal cost per token is the rental rate divided by throughput, and throughput, which depends on model size and serving efficiency that closed labs do not disclose, moves far more than rental rates do. Between early and mid 2024 the price of a given level of measured capability fell by more than ninety-nine percent while H100 rental rates rose (Appendix D.6). A regression of closed-model prices on rental rates would measure the efficiency trend, not demand curvature.

Open-weight models allow a cleaner test. Their developers sell them through their own APIs, and the same weights are served by several independent hosts, whose prices are a market-based proxy for marginal cost. Across 43 such models, the elasticity of the developer's price with respect to the host price is 0.64 to 0.81 in ordinary least squares, but the host price measures marginal cost with error, which biases that estimate toward zero. A split-sample instrument that corrects for this gives 0.92, with a 95 percent interval of 0.79 to 1.15, so constant markups, $\rho = 1$, cannot be rejected. On this evidence the domestic sign survives comfortably, and the importer's rate is the fragile number, between about 6 and 20 percent.

The caveat is substantial. Developers price their open-weight models at almost exactly the hosts' level, because anyone can serve the same weights, so this is a nearly competitive market, and in a nearly competitive market pass-through close to one is expected whatever the curvature of demand. The test validates the method and bounds what open-weight pricing implies; it does not measure pass-through for the closed frontier models whose markups drive the market-power term. That number remains unmeasured, and Appendix D.6 sets out the data, the specifications and the limitations in full.

Two things follow. First, the sign of the optimal usage tax is not settled by the model alone; it depends on whether ρ is above or below one, a quantity that has not

Table 5: Optimal usage tax by pass-through, entry blocked

Value of ρ	Source	Domestic $t_U^\star$	Importer $t_U^\star$
1.00	Constant elasticity (model)	−15.5%	+6.4%
0.92	Open-weight models, IV point estimate	−14.0%	+10.9%
0.79	Open-weight models, IV lower bound	−10.9%	+20.0%
0.64	Open-weight models, lowest OLS estimate	−5.7%	+35.1%
0.53	Domestic sign threshold	0	+52%

Notes. First-order values from (10) at the central calibration: $\mu = 1.5$, $(\Omega + D)/p = 6.44\%$, $g_S = 0.66$, $\sigma_q = 1.965$. The open-weight estimates come from a market in which developers price at roughly the level of third-party hosts and are not estimates of pass-through for closed frontier models (Appendix D.6). Because ρ changes as a tax grows, first-order values overstate the importer's rate when it is large.

been measured for the closed frontier models whose markups matter. Under the demand system the model generates, the domestic optimum is negative; under demand curves that make providers absorb a substantial share of a tax, it is positive, and for an importer it can be large. Second, what survives across demand systems is the ordering by ownership: the importer's optimal rate exceeds the domestic one in each, because rent that accrues abroad is not part of domestic welfare. Appendix D.5 also shows that ρ can be measured without any tax variation, and Appendix D.6 reports what public data can and cannot say about it.

6.6 Who pays, and who benefits

The paper has so far kept efficiency and distribution in separate compartments. But the missing-market wedge γ has a distributional character of its own, and once it is made explicit the environmental and the equity cases for taxing at the meter stop being separate arguments and become the same argument.

Recall what γ is: the part of the capacity and transmission cost caused by a data center that is recovered from all ratepayers rather than from the facility. It is collected through ordinary electricity bills, so households bear it roughly in proportion to their electricity consumption. Consumption rises with income but far less than proportionately, so the charge is steeply regressive as a share of income. The median U.S. household spends about 2.9 percent of income on energy, low-income households between 6 and 15 percent, and a quarter of low-income households above 15 percent (Drehobl et al., 2020; Rocky Mountain Institute, 2026). A socialized grid cost is therefore borne, relative to income, something like three to five times as heavily by the poorest households as by the median one.

The other side of the ledger runs the opposite way. The benefits of the buildout—rents

captured by providers and surplus enjoyed by users of AI—accrue disproportionately to the affluent: to shareholders of a small number of firms, and to professional and business users who are the early adopters of paid services. The socialization of γ is thus a transfer up the income distribution, with a large and growing base.

The consequence for policy is that the welfare weight on the grid-cost term is doing real work, and it points the same way as the efficiency argument. Charging the data center for the capacity it causes corrects the externality. Whether it also stops the regressive transfer depends on the instrument: a dedicated large-load rate class reallocates the cost to the facility and so removes the household cross-subsidy, whereas a load-based charge collected by the treasury corrects the incentive but leaves the utility's cost allocation in place unless its revenue is used to offset household bills. The token tax does neither well: it reaches the grid cost only through the scale of AI use, leaves the utility's cost allocation and the facility's energy intensity untouched, and its revenue is recycled with no particular distributional tilt. It is unusual for a Pigouvian instrument to have its efficiency and equity cases coincide—the standard worry about carbon taxes is that they are regressive in use even when efficient (Hassett et al., 2009). Here the regressivity is in the status quo of socialized cost, and pricing the externality removes it rather than creating it.[21]

7 Policy assessment

The preceding sections characterize an optimum. A finance ministry choosing among live proposals needs something narrower: what each would raise, what it would cost, who would bear it, and whether it can be administered. This section puts the framework to that use. We score stylized versions of the four instruments under discussion, express the comparison in a common welfare metric, ask which of them can be administered, and then turn to why the political system may nonetheless reach for the worst one.

7.1 A common metric: the marginal value of public funds

Revenue and deadweight loss are not comparable across instruments whose burdens fall on different people, and a deadweight-loss triangle misses the first-order interaction with

[21] Two qualifications. The incidence of the benefit side is inferred from adoption patterns rather than measured, so the transfer is regressive with confidence on the cost side and less on the benefit side; and if a large load lowers average rates by improving asset utilization, the sign can flip. Both point to measuring γ rather than assuming it, and neither disturbs the central point that the grid-cost term carries a distributional charge the token tax cannot touch.

Table 6: Scoring stylized proposals with full welfare accounting

			Welfare change (% GDP)		
		Revenue	Entry blocked		Free entry
Instrument	Rate	(% GDP)	domestic	importer	any
Token / usage tax	10%	0.124	−0.052	+0.004	−0.087
DST-style levy on AI revenue	3%	0.042	−0.014	+0.004	−0.022
Data center emissions charge	d per kWh	0.041	−0.029	−0.015	−0.035
Cash flow tax on rents	50%	0.250	$0^{\dagger}$	$0^{\dagger}$	no base

Notes. Central calibration, $\mu = 1.5$, $(\Omega + D)/p = 6.44\%$, $g_S = 0.66$, $\theta = 1.5\%$, constant-elasticity demand. Each entry is the change in domestic welfare from introducing the instrument alone, computed from the welfare functions of Propositions 2, 7 and 8: it counts the pre-existing markup, the externality and the wage effects as well as the tax, not only a Harberger triangle. Revenue is after behavioral response, with entry blocked. The emissions charge is scored on the scale margin only, holding energy intensity and training electricity fixed; those are the margins on which it has an advantage (Proposition 3, Appendices B.4 and B.5), so its entry here is a lower bound. Under constant-elasticity demand it is passed through at μ, so each unit priced also transfers $(\mu - 1)D$ to providers. †The cash flow tax is neutral on an ideal base; §7.2 puts a magnitude on the departure. Under free entry there is no rent for it to reach. Section 6.5 reports how the entries change when providers absorb part of a tax.

the distortions already present. Table 6 therefore reports the full change in domestic welfare from each instrument, and dividing by revenue gives a net cost per dollar raised in the spirit of the marginal value of public funds of Hendren and Sprung-Keyser (2020).

Three rankings follow. The cash flow tax is cheapest wherever it has a base: with entry blocked it raises about 0.25 percent of GDP at no first-order cost, but under free entry there is no rent for it to reach. The usage tax and the DST lower domestic welfare in both regimes, by 0.42 to 0.75 dollars per dollar raised for the usage tax; for an importer with entry blocked they are roughly neutral, consistent with an optimal rate of about six percent. The emissions charge, scored on the scale margin alone, also lowers welfare, because under constant-elasticity demand it is passed through at μ and transfers $(\mu-1)D$ per unit to providers. With energy efficiency held fixed, correcting the externality on the usage side does at least as well: the usage tax and the charge act on the same margin, and the charge's distinctive advantages, the efficiency margin and the training electricity a token tax never reaches, are exactly what this static scoring leaves out.

The case for the meter therefore rests on those margins and not on the scale margin, and it has to be made alongside a correction for market power, since pricing the kilowatt-hour in a marked-up sector passes through with amplification. Two further qualifications run against the charge. Its efficiency advantage is a long-run one, opening as durable investment responds. And the corrective benefit must be netted of relocation, since Burke and Abayasekara (2018) find the large long-run industrial elasticity arises precisely because electricity-intensive activity clusters in low-price states, which through our model

is leakage rather than abatement.

7.2 The zeros are not honest zeros

Proposition 4 delivers exact neutrality, and Table 6 therefore carries zeros for the cash flow tax. We do not believe those zeros and neither should a finance ministry. They are a property of an idealized base—immediate expensing of all investment, full loss offset carried forward with interest, no distinction between normal return and rent, no ability to relocate the taxable entity—and no enacted statute has had all four. Appendix B.11 lists the four channels through which a wedge creeps back in, each with its drafting remedy.

The claim we defend is weaker and still decisive: the distortion is smaller than a usage tax's by an order of magnitude, not nil, and the magnitudes are second order in the residual wedge, which is what makes the comparison robust even though the zero is not. Suppose the channels together leave an effective wedge on the marginal return of a tenth of the statutory rate, and an investment elasticity of unity—both pessimistic. On the 0.5 percent of GDP of rent at the central markup the deadweight loss is $\frac{1}{2}(0.05)^2 \times 0.5 \approx 0.0006$ percent of GDP, or roughly *a quarter of a cent per dollar raised*, against 42 cents for the usage tax and 34 for the DST in Table 6. The ordering survives any plausible value, for a structural reason: the usage tax deepens an existing underprovision, while the cash flow tax distorts only whatever fraction of the normal return the statute fails to expense. Closing the gap is a drafting problem, not an economic one.

Administrability. Optimal tax formulae assume the base is observable, and following Keen and Slemrod (2017) we treat administrability as part of instrument choice. The emissions charge is the easy case: the base is metered by the utility at the point of delivery for billing purposes that already exist, tons follow from kilowatt-hours and a published grid factor, and electricity delivered to a physical site cannot be booked in another jurisdiction. The rent tax is demanding but familiar, and its failure mode is political rather than technical. The usage tax is trivially easy to administer, and that is part of the problem: administrative convenience is precisely why it attracts legislatures, and a low enforcement cost is a genuine advantage that must be weighed rather than dismissed—our claim is that it does not come close to offsetting a base that is wrong by an order of magnitude on the harm-to-value ratio. The border adjustment is the hard case and we do not overclaim it; Appendix E.1 sets out why, and the three fallbacks in descending order of fidelity.

7.3 Why we might get the wrong instrument

If the usage tax is poorly suited to every stated goal, why is it the instrument most often proposed? The model contains the beginning of an answer, and it comes from the entry result rather than from any appeal to confusion among policymakers.

Remark 3 (Incumbents and the usage tax in transition). In neither steady state of Section 5.6 do incumbents gain from a usage tax: with entry blocked it passes through at one and lowers their profit, and with free entry they earn nothing either way. The preference can arise in transition. Suppose incumbents have sunk their fixed costs while the number of firms in the next model generation is set by free entry. A cash flow tax raising revenue R lowers incumbent profit by R. A usage tax raising the same revenue lowers the number of firms the next generation sustains (Proposition 6) and so raises the markup on the sales incumbents retain. If incumbents value that future margin enough, they can prefer the usage tax to the revenue-equivalent cash flow tax, the reverse of the planner's ranking.

The content is a positive prediction, not a normative claim: the instrument that is worst for welfare is the one the most concentrated and best-organized interest has least reason to oppose, because it raises rivals' entry cost and lifts margins on the sales that remain. This yields a testable implication that distinguishes the account from mere pessimism. If incumbents prefer the usage tax for the entry-barrier reason, and if firms differ in size and sunk cost, which our symmetric model does not capture, their revealed positions should track those differences: the largest labs least opposed to usage levies and most opposed to rent taxes and interoperability mandates, smaller entrants holding the reverse ranking. That is a pattern one can look for in lobbying disclosures. Appendix E.2 develops two further forces that push the same way and the normative implication, which is not defeatism but sequence: because the worst instrument is the path of least resistance, the reforms most worth defending are the ones the political system will fight hardest.

8 Conclusion

The debate about taxing AI has run four different arguments together. Once separated, they point to different instruments and, in one case, in the opposite direction to the one usually assumed.

The externality is real, but it lives in the electricity system rather than in the token. The residual wedge there is about sixty percent of the price of a kilowatt-hour at the

central damage estimate and more than the whole of it at the top of the range, driven mainly by unpriced carbon, against a few percent for a tax on AI usage—and no single ad valorem rate can track a harm whose ratio to value varies by an order of magnitude across workloads, whereas a charge per kilowatt-hour tracks it by construction. Pricing at the meter also reaches the efficiency margin, so its advantage grows with the horizon, and its efficiency case coincides with its equity case, since the socialized grid cost it corrects is itself a regressive transfer.

The redistributive argument is real and small. It is smaller than the robot-tax literature implies because that literature identifies the losing group with the exposed occupation, which is a modelling convention rather than a fact, and because the parameter usually read as mobility is the same parameter that makes the losing group identifiable in the first place. The market-power argument runs the other way and, under the demand system the model generates, is the largest of the three. With entry blocked it is the shareholder-weighted Lerner index; under free entry it reflects the markup a tax creates by driving out entrants.

Should policy compensate the losers from technological change at all? The tradition we build on answers yes, but through the general tax and transfer system, and for the reason that drives this paper. An instrument aimed at a technology must predict who will lose, whereas an income-tested transfer conditions on who has lost. When both the sign and the size of the distributional effect are uncertain, as Table 2 shows they are for AI, that difference is decisive: a technology tax calibrated to the wrong sign redistributes toward the better-off, and the measured range of Δ_ℓ contains both signs. The guidance that survives this uncertainty is to compensate through instruments that condition on realized outcomes, such as an earned income tax credit or targeted adjustment assistance, to price the technology only for the external costs it imposes, and to measure the exposure gap before legislating on it. This is also what a fully specified model of automation delivers: Kleven and Zidar (2026) find that automation calls for a more progressive labor tax, with a work subsidy at the bottom roughly twice as generous as without automation, rather than a tax on the machines themselves.

Netting the three, and imposing the restrictions the model generates—mobility tied to the exposure gap, markups within the Lerner bound, the shareholder weight tied to the weight on routine workers—the optimal usage tax on a domestic sector is negative in 97 percent of the parameter space and above five percent in under one percent of it. We do not recommend a subsidy: underprovision is a problem for entry, interoperability and procurement policy, and if those instruments address it the baseline wedge on usage

is about 6 percent. The sign rests on pass-through, which no one has measured for AI: where providers absorb part of a tax, the tax extracts rent and the optimum turns positive. What is robust is the comparison by ownership, and the magnitude: under the model's demand system the instrument being legislated sits at or above the top of what the model supports.

Each of the corrections removes an assumption rather than adding evidence. What survives all of them is a division of labor among instruments. The electricity wedge is large, and only a charge at the meter reaches the efficiency margin and the electricity used in training; on the scale margin alone it does no better than a usage tax, because a marked-up sector passes it through with amplification, so it belongs alongside competition policy rather than in place of it. Rent is reached by a tax on rent, not on transactions. And the case for taxing tokens, if there is one, rests on a pass-through rate that should be measured before it is legislated.

References

Acemoglu, D., Manera, A., and Restrepo, P. (2020). Does the US tax code favor automation? *Brookings Papers on Economic Activity*, 51(1):231–300.

Ales, L., Kurnaz, M., and Sleet, C. (2015). Technical change, wage inequality, and taxes. *American Economic Review*, 105(10):3061–3101.

Atkinson, A. B. and Stiglitz, J. E. (1976). The design of tax structure: Direct versus indirect taxation. *Journal of Public Economics*, 6(1–2):55–75.

Auerbach, A. J. and Devereux, M. P. (2018). Cash-flow taxes in an international setting. *American Economic Journal: Economic Policy*, 10(3):69–94.

Auerbach, A. J. and Hines Jr., J. R. (2001). Perfect taxation with imperfect competition. Working Paper 8138, National Bureau of Economic Research.

Autor, D. (2024). Applying AI to rebuild middle class jobs. Working Paper 32140, National Bureau of Economic Research. Published in *NOEMA Magazine* as "How AI Could Help Rebuild The Middle Class".

Autor, D. and Thompson, N. (2025). Expertise. *Journal of the European Economic Association*, 23(4):1203–1271.

Autor, D. H. and Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the US labor market. *American Economic Review*, 103(5):1553–1597.

Autor, D. H., Dorn, D., and Hanson, G. H. (2013). The China syndrome: Local labor market effects of import competition in the United States. *American Economic Review*, 103(6):2121–2168.

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

Bick, A., Blandin, A., Deming, D. J., and Schumacher, T. (2026). What work does generative AI do? Working Paper 2026-017, Federal Reserve Bank of St. Louis. Task- and occupation-level generative-AI adoption indexes from the Real-Time Population Survey. The replication package containing the indexes had not been released at the time of writing.

Bonney, K., Breaux, C., Dinlersoz, E., Foster, L., Haltiwanger, J., and Pande, A. (2026). The microstructure of AI diffusion: Evidence from firms, business functions, and worker tasks. CES Working Paper

26-25, U.S. Census Bureau, Center for Economic Studies. Based on the 2026 AI supplement to the Business Trends and Outlook Survey (BTOS).

Borenstein, S. (2025). What will data centers do to your electric bill? Energy Institute at Haas, Blog Post.

Borenstein, S. and Bushnell, J. B. (2022). Do two electricity pricing wrongs make a right? cost recovery, externalities, and efficiency. *American Economic Journal: Economic Policy*, 14(4):80–110.

Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at work. *Quarterly Journal of Economics*, 140(2):889–942.

Burke, P. J. and Abayasekara, A. (2018). The price elasticity of electricity demand in the United States: A three-dimensional analysis. *The Energy Journal*, 39(2):123–146.

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

Cournot, A. A. (1838). *Recherches sur les principes mathématiques de la théorie des richesses*. Hachette, Paris.

De Loecker, J., Eeckhout, J., and Unger, G. (2020). The rise of market power and the macroeconomic implications. *Quarterly Journal of Economics*, 135(2):561–644.

de Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule*, 7(10):2191–2194.

Delipalla, S. and Keen, M. (1992). The comparison between ad valorem and specific taxation under imperfect competition. *Journal of Public Economics*, 49(3):351–367.

Diamond, P. A. and Mirrlees, J. A. (1971). Optimal taxation and public production I: Production efficiency. *American Economic Review*, 61(1):8–27.

Dorn, D. (2009). Essays on inequality, spatial interaction, and the demand for skills. *Dissertation 3613, University of St. Gallen*. Source of the `occ1990dd` harmonized occupation scheme and the associated labor-supply weights; files at `https://www.ddorn.net/data.htm`.

Drehobl, A., Ross, L., and Ayala, R. (2020). How high are household energy burdens? an assessment of national and metropolitan energy burden across the united states. Technical report, American Council for an Energy-Efficient Economy.

Eisfeldt, A. L., Schubert, G., Taska, B., and Zhang, M. B. (2023). Generative ai and firm values. *NBER Working Paper 31222*. Occupation-level generative-AI exposure scores `genaiexp_estz` by SOC 2010.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308. Occupation-level α, β, ζ exposure scores under both the human and the GPT-4 rubric.

Energy and Environmental Economics (E3) (2026). Understanding the drivers of rising electricity rates and the role of data centers. Technical report, E3 White Paper.

Felten, E. W., Raj, M., and Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 42(12):2195–2217. AI Occupational Exposure (AIOE) scores by SOC 2010, Data Appendix A.

Fischer, C. and Fox, A. K. (2012). Comparing policies to combat emissions leakage: Border carbon adjustments versus rebates. *Journal of Environmental Economics and Management*, 64(2):199–216.

Galle, S., Rodríguez-Clare, A., and Yi, M. (2023). Slicing the pie: Quantifying the aggregate and distributional effects of trade. *The Review of Economic Studies*, 90(1):331–375.

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

Hassett, K. A., Mathur, A., and Metcalf, G. E. (2009). The incidence of a U.S. carbon tax: A lifetime and regional analysis. *The Energy Journal*, 30(2):155–177.

Hendren, N. and Sprung-Keyser, B. (2020). A unified welfare analysis of government policies. *The Quarterly Journal of Economics*, 135(3):1209–1318.

Hoel, M. (1996). Should a carbon tax be differentiated across sectors? *Journal of Public Economics*, 59(1):17–32.

Hsieh, C.-T., Hurst, E., Jones, C. I., and Klenow, P. J. (2019). The allocation of talent and U.S. economic growth. *Econometrica*, 87(5):1439–1474.

Ide, E. and Talamàs, E. (2025). Artificial intelligence in the knowledge economy. *Journal of Political Economy*, 133(12):3761–3800.

International Energy Agency (2025). Energy and AI. Technical report, IEA, Paris.

International Energy Agency (2026). Key questions on energy and AI. Technical report, IEA, Paris.

Jacobs, M. (2025). Data centers and the connection costs loophole: Billions in transmission costs passed to PJM customers. Technical report, Union of Concerned Scientists.

Joint Legislative Audit and Review Commission (2024). Data centers in Virginia. Technical report, Commonwealth of Virginia.

Kaplow, L. (2012). Optimal control of externalities in the presence of income taxation. *International Economic Review*, 53(2):487–509.

Keen, M. (1998). The balance between specific and ad valorem taxation. *Fiscal Studies*, 19(1):1–37.

Keen, M. and Slemrod, J. (2017). Optimal tax administration. *Journal of Public Economics*, 152:133–142.

Kleven, H. and Zidar, O. M. (2026). Automation and optimal taxation: A task-based theory. Working Paper 35747, National Bureau of Economic Research.

Korinek, A. and Lockwood, L. M. (2025). Public finance in the age of AI: A primer. Working paper, Brookings Center on Regulation and Markets.

Kydland, F. E. and Prescott, E. C. (1977). Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy*, 85(3):473–491.

Labandeira, X., Labeaga, J. M., and López-Otero, X. (2017). A meta-analysis on the price elasticity of energy demand. *Energy Policy*, 102:549–568.

Löbbing, J. (2024). Redistributive income taxation with directed technical change. *Journal of Public Economics*.

Luccioni, A. S., Jernite, Y., and Strubell, E. (2024). Power hungry processing: Watts driving the cost of AI deployment? In *ACM Conference on Fairness, Accountability, and Transparency*, pages 85–99.

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

Markusen, J. R. (1975). International externalities and optimal tax structures. *Journal of International Economics*, 5(1):15–29.

Monitoring Analytics (2025). Analysis of the 2025/2026 RPM base residual auction. Technical report, Independent Market Monitor for PJM.

Myles, G. D. (1996). Imperfect competition and the optimal combination of ad valorem and specific taxation. *International Tax and Public Finance*, 3(1):29–44.

Naito, H. (1999). Re-examination of uniform commodity taxes under a non-linear income tax system and its implication for production efficiency. *Journal of Public Economics*, 71(2):165–188.

National Center for O*NET Development (2026). O*net 30.3 database. U.S. Department of Labor, Employment and Training Administration. `https://www.onetcenter.org/database.html`. Used here: `education.csv` (Required Level of Education, scale RL), `job_zones.csv`, `occupation_data.csv`. Licensed CC BY 4.0.

Natural Resources Defense Council (2025). Data center growth and the cost to PJM ratepayers. Technical report, NRDC Issue Brief.

Nordhaus, W. D. (2017). Revisiting the social cost of carbon. *Proceedings of the National Academy of Sciences*, 114(7):1518–1523.

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

O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., and Dafoe, A. (2020). The windfall clause: Distributing the benefits of AI for the common good. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*, pages 327–331.

Riu, I., Patel, K., Mettetal, L., et al. (2025). Tailored for scale: Designing electric rates and tariffs for large loads. Technical report, Energy and Environmental Economics (E3).

Rocky Mountain Institute (2026). Energy burden across income groups: A fifty-state analysis. Technical report, RMI.

Rothschild, C. and Scheuer, F. (2013). Redistributive taxation in the Roy model. *The Quarterly Journal of Economics*, 128(2):623–668.

Sachs, D., Tsyvinski, A., and Werquin, N. (2020). Nonlinear tax incidence and optimal taxation in general equilibrium. *Econometrica*, 88(2):469–493.

Saez, E. (2004). Direct or indirect tax instruments for redistribution: Short-run versus long-run. *Journal of Public Economics*, 88(3–4):503–518.

Saez, E., Schoefer, B., and Seim, D. (2019). Payroll taxes, firm behavior, and rent sharing: Evidence from a young workers' tax cut in Sweden. *American Economic Review*, 109(5):1717–1763.

Sandmo, A. (1975). Optimal taxation in the presence of externalities. *Swedish Journal of Economics*, 77(1):86–98.

Scheuer, F. (2014). Entrepreneurial taxation with endogenous entry. *American Economic Journal: Economic Policy*, 6(2):126–163.

Scheuer, F. and Werning, I. (2016). Mirrlees meets Diamond–Mirrlees. NBER Working Paper 22076, National Bureau of Economic Research.

Scheuer, F. and Werning, I. (2017). The taxation of superstars. *Quarterly Journal of Economics*, 132(1):211–270.

Shehabi, A., Smith, S. J., Hubbard, A., et al. (2024). 2024 united states data center energy usage report. Technical report, Lawrence Berkeley National Laboratory.

Stiglitz, J. E. (1982). Self-selection and Pareto efficient taxation. *Journal of Public Economics*, 17(2):213–240.

Suits, D. B. and Musgrave, R. A. (1953). Ad valorem and unit taxes compared. *Quarterly Journal of Economics*, 67(4):598–604.

Thümmel, U. (2023). Optimal taxation of robots. *Journal of the European Economic Association*, 21(3):1154–1190.

Tomlinson, K., Jaffe, S., Wang, W., Counts, S., and Suri, S. (2025). Working with AI: Measuring the applicability of generative AI to occupations. *arXiv preprint arXiv:2507.07935*. Microsoft Research. Version 6, 22 December 2025. Occupation-level AI applicability scores derived from 200,000 Bing Copilot conversations classified into O*NET work activities; aggregated measures at `https://github.com/microsoft/working-with-ai`.

U.S. Bureau of Labor Statistics (2018). 2010 to 2018 SOC crosswalk. Standard Occupational Classification. `https://www.bls.gov/soc/2018/soc_2010_to_2018_crosswalk.xlsx`.

U.S. Environmental Protection Agency (2023). Report on the social cost of greenhouse gases: Estimates incorporating recent scientific advances. Technical report, EPA.

Webb, M. (2020). The impact of artificial intelligence on the labor market. *SSRN Working Paper 3482150*. Occupation-level software, robot and AI exposure percentiles by `occ1990dd`, with 2010 labor-supply weights; replication file `exposure_by_occ1990dd_lswt2010.csv`.

Wiser, R., Barbose, G., Gorman, W., et al. (2026). Retail electricity price trends and drivers: Data update, 2026 edition. Technical report, Lawrence Berkeley National Laboratory.

Yin, M., Vu, H., and Persico, C. (2026). How (un)Stable are LLM occupational exposure scores? evidence from multi-model replication. Working Paper 35110, National Bureau of Economic Research, Cambridge, MA. Author forenames not given in the source we transcribed this entry from; complete before submission.

A Proofs

Proof of Lemma 1

Unit cost of the final good is $c_Y = c_A^{\lambda} w_N^{1-\lambda}/\Lambda$ by (1), and by (2) $\partial \ln c_A / \partial \ln q = s$, so at given factor prices $\mathrm{d}\ln c_Y = \lambda s \,\mathrm{d}\ln q = \theta \,\mathrm{d}\ln q$. By Shephard's lemma this is the first-order loss of real income at given factor prices. It is not, in general, the equilibrium change in gross output, $\mathrm{d}\ln Y = \lambda(1-s)\mathrm{d}\ln L_R + (1-\lambda)\mathrm{d}\ln L_N + \theta\,\mathrm{d}\ln a$, which depends on the quantity responses; the welfare calculations of Section 4 use the money-metric statement only. ■

Proof of Lemma 2 and Proposition 5

Relative wages. From (6), $w_R/w_N = [\lambda(1-s)/(1-\lambda)]\cdot(L_N/L_R)$, so with $x \equiv \mathrm{d}\ln(w_R/w_N)$ and using (2),

$$x \;=\; s(\sigma-1)\,\mathrm{d}\ln q \;-\; \mathrm{d}\ln(L_R/L_N).$$

Relative task labor responds through two margins. The intensive hours margin contributes $\varepsilon_L x$. The reallocation margin contributes $\mathcal{E}x$, where $\mathcal{E}$ is defined by $\mathrm{d}\ln(L_R/L_N)\big|_{\text{reallocation}} = \mathcal{E}\,x$. The two are additive because they operate on distinct decisions: how many hours to supply, and to which task. Substituting and solving gives $x = s(\sigma-1)\mathrm{d}\ln q/(1+\varepsilon_L+\mathcal{E})$.

The reallocation identity. Under Fréchet sorting the average efficiency of group-i workers in task R is proportional to $\ell_{iR}^{-1/\kappa}$, so group i's effective routine supply is proportional to $\ell_{iR}^{1-1/\kappa}$ and responds as $\mathrm{d}\ln L_{iR} = (\kappa-1)(1-\ell_{iR})\,x$, with $\mathrm{d}\ln L_{iN} = -(\kappa-1)\ell_{iR}\,x$. Aggregating over groups with employment shares π and $1-\pi$ and routine share $h = \pi\ell_{1R} + (1-\pi)\ell_{2R}$,

$$\mathcal{E} = \frac{\kappa-1}{h(1-h)}\sum_i \pi_i\,\ell_{iR}(1-\ell_{iR}) = \frac{\kappa-1}{h(1-h)}\Big[h(1-h) - \pi(1-\pi)\Delta_\ell^2\Big],$$

using $\sum_i \pi_i \ell_{iR}^2 = h^2 + \pi(1-\pi)\Delta_\ell^2$. This is the expression in Lemma 2. In the symmetric specification of (5), with equal group masses and $A_{1R}/A_{1N} = e^{\varrho/2}$, $A_{2R}/A_{2N} = e^{-\varrho/2}$, evaluated at $w_R = w_N$, we have $\ell_{2R} = 1-\ell_{1R}$ and $h = \frac{1}{2}$, so $\mathcal{E} = (\kappa-1)(1-\Delta_\ell^2)$, and $\ell_{1R} = (1+e^{-\kappa\varrho/2})^{-1}$ gives $\Delta_\ell = \tanh(\kappa\varrho/4)$. Away from that point the exposure gap depends on relative wages, and the calibration uses the general expression.

The distributional value. By (B.4), $\mathrm{d}\ln(W_1/W_2) = \Delta_\ell\,x$: the switching margin drops out by the envelope theorem, since the marginal switcher is indifferent, and the relative welfare movement between groups is the relative task-wage movement scaled by the ex-

posure gap. Substituting into the evaluation of Ω in the proof of Proposition 2, with $\hat{g}_1$ defined on ex ante groups so that $\sum_i \hat{g}_i W_i = 0$ holds throughout the reform, replaces $s(\sigma-1)/(1+\varepsilon_L)$ by $\Delta_\ell\, s(\sigma-1)/(1+\varepsilon_L+\mathcal{E})$, with the income base taken to be group earnings as in that proof, and leaves every other step unchanged, since neither Δ_ℓ nor $\mathcal{E}$ enters the resource cost of a unit of AI or the external cost D.

Comparative statics. Change variable to $u \equiv \kappa\varrho/4 \geq \varrho/4$, so that $\Delta_\ell = \tanh u \equiv t$ and, with $\operatorname{sech}^2 u = 1 - t^2$,

$$\mathcal{E} \;=\; (\kappa-1)(1-t^2) \;=\; \Big(\frac{4u}{\varrho} - 1\Big)\operatorname{sech}^2 u.$$

For (i): $\mathcal{E} = 0$ at $\kappa = 1$, $\mathcal{E} \to 0$ as $u \to \infty$ because $u \operatorname{sech}^2 u \leq 4u\, e^{-2u} \to 0$, and $\mathcal{E} > 0$ in between. Part (ii) is immediate from the strict monotonicity of tanh.

For (iii), write $A \equiv 1 + \varepsilon_L$ and $f(u) \equiv t/(A+\mathcal{E})$, so that Ω/p is $f(u)$ times a positive constant. Using $d\mathcal{E}/du = \operatorname{sech}^2 u\,[\,4/\varrho - 2t(\kappa-1)]$,

$$f'(u) \;=\; \frac{\operatorname{sech}^2 u}{(A+\mathcal{E})^2}\Big[A + (\kappa-1)(1+t^2) - \frac{4t}{\varrho}\Big].$$

Since $4/\varrho = \kappa/u$ and $t \leq u$, the last term satisfies $4t/\varrho = \kappa\, t/u \leq \kappa$, so the bracket is at least $A - 1 + (\kappa-1)t^2 = \varepsilon_L + (\kappa-1)t^2 > 0$. Hence $f'(u) > 0$ for all $\kappa > 1$, uniformly in ϱ, and $f \to 1/A$, the fixed-assignment value, as $u \to \infty$.

Part (iv) follows from $\mathcal{E} \geq 0$, which gives $\Omega/p \leq \Delta_\ell \cdot (1-\tau)\hat{g}_1(\sigma-1)(1-s)/[(1+\varepsilon_L)\sigma_q]$, with equality only when $\mathcal{E} = 0$, that is at $\kappa = 1$ or in the limit $\kappa \to \infty$. □

Proof of Proposition 1

Under these conditions the economy satisfies the hypotheses of Diamond and Mirrlees (1971). Producer prices can be held equal to social marginal costs while any feasible revenue is raised on final uses and factor incomes. A positive t_U places a wedge on the intermediate margin, which by Lemma 1 raises the unit cost of the economy by $\lambda s\, t_U$ to first order and lowers output. Replace t_U by zero and adjust the income tax to hold revenue constant: the tagged income tax attains the same distribution at the same revenue with strictly larger output. Separability and type-contingent taxes also make AI intensity uninformative once earnings are observed, so by Atkinson and Stiglitz (1976) no differentiation of labor taxes by AI usage is desirable. ■

Proof of Proposition 2

Hold the number of providers fixed. With constant-elasticity demand and symmetric providers the markup μ does not depend on the tax, so the producer price $p = \mu m$ is fixed and the buyer price $q = (1 + t_U)p$ moves one for one with the tax. Weighting buyer surplus and public funds at one, which Lemma 3 justifies for the common component of income, domestic welfare changes by

$$dW \;=\; -a\,dq \;+\; d(t_U\,p\,a) \;+\; (1-f)(p-m)\,da \;-\; D\,da \;+\; dW^{\text{dist}},$$

the terms being buyer surplus, tax revenue, domestically owned provider profit, the externality, and the distributional value of wage movements. Since $dq = p\,dt_U$, the first term cancels the mechanical part of the second, leaving $dW = [\,t_U p + (1-f)(p-m) - D\,]\,da + dW^{\text{dist}}$. Write $dW^{\text{dist}} = -\Omega\,da$, where Ω is the distributional value per unit of AI forgone. Setting $dW = 0$,

$$t_U^{\star} \;=\; \frac{\Omega + D}{p} \;-\; (1-f)\Big(1 - \frac{1}{\mu}\Big).$$

The bracket is linear in t_U, so this holds at any rate, not only locally. The fixed cost does not appear because, with the number of providers fixed, it does not vary with a.

To evaluate Ω: with the income tax optimal, the distributional value of the reform is the sum of net welfare weights times changes in net income, and by Lemma 3 the common component of the wage change has no first-order distributional value. Let B_i be the gross earnings of ex ante group i and $\hat{g}_i$ its excess welfare weight per dollar, normalized so that $\sum_i \hat{g}_i B_i = 0$. Under Fréchet sorting the share of group i's earnings that comes from routine work equals its routine employment share ℓ_{iR}, so $\mathrm{d}\ln(B_1/B_2) = \Delta_\ell\,x$ with $x = \mathrm{d}\ln(w_R/w_N)$, and

$$\frac{dW^{\text{dist}}}{\mathrm{d}\ln q} \;=\; (1-\tau)\,\hat{g}_1 B_1\,\Delta_\ell\,\frac{x}{\mathrm{d}\ln q} \;=\; (1-\tau)\,\hat{g}_1 B_1\,\Delta_\ell\,\frac{s(\sigma-1)}{1+\varepsilon_L+\mathcal{E}}$$

by Lemma 2. The associated fall in AI use is $-da = \sigma_q a\,\mathrm{d}\ln q$ by (7), so

$$\frac{\Omega}{p} \;=\; \frac{(1-\tau)\hat{g}_1\Delta_\ell\,s(\sigma-1)}{(1+\varepsilon_L+\mathcal{E})\,\sigma_q}\cdot\frac{B_1}{p\,a}.$$

By (6), $p\,a = \lambda s Y$, so $B_1/(pa) = \beta_1/s$ with $\beta_1 \equiv B_1/(\lambda Y)$, and s cancels. Under fixed assignment group 1 is the routine workforce, $B_1 = w_R L_R = \lambda(1-s)Y$ and $\beta_1 = 1-s$, which recovers the two-type formula. ■

Proof of Proposition 3

Under t_E the provider's energy price becomes $p_e + t_E$ and the efficiency condition becomes $\varphi'(\xi) = -(p_e + t_E)e'(\xi)$. At $t_E = d$ the provider internalizes the full social cost of a kilowatt-hour, chooses $\xi^\star$, and passes the resulting marginal cost into the price, so the scale margin is correct as well. Under t_U the provider's energy price is unchanged, the efficiency condition is unchanged, and $\xi = \xi^0$ regardless of the rate: a usage tax is a wedge on output, not on energy, and the energy intensity of provision is not a function of the output price. The residual loss is the excess of the social cost of provision at ξ^0 over its minimized value at $\xi^\star$, which is Δ. Strict convexity makes $\Delta > 0$ unless $e' \equiv 0$. ■

Proof of Proposition 4

The provider maximizes $(1 - z)[P(Q)Q - m(\xi)Q - F]$. Since z scales the objective, the maximizing Q, P, and ξ are unchanged, and revenue $z\Pi$ is raised without distortion. The usage tax raises $t_U pa$ and induces $\mathrm{d}\ln a = -\sigma_q t_U$ to first order, generating a Harberger loss of order $\frac{1}{2}\sigma_q\, pa\, t_U^2 > 0$. ■

Proof of Proposition B.1

(i) A marginal unit of domestic abatement reduces global emissions by $1-\ell$, so its marginal social benefit is $(1 - \ell)d$ per kilowatt-hour while its marginal cost is borne domestically at full value. This is the leakage adjustment of Markusen (1975) and Hoel (1996). (ii) AI consumed domestically but served from abroad escapes an origin charge; its embodied residual damage is D per unit. A levy on the buyer at D/p on those services, credited for any charge paid at origin, applies exactly that wedge and does not depend on where the server sits. (iii) With every unit of domestic consumption bearing D wherever it is served, relocating compute no longer reduces the charge on that consumption, so the leakage rate relevant to (i) is zero for it and the origin charge can be set at d. Leakage through production for foreign markets is unaffected. Fischer and Fox (2012) compare border carbon adjustments with output-based rebates numerically for goods trade, and the leakage logic here is the same. ■

Proof of Proposition 6

With constant-elasticity demand $Q = Aq^{-\sigma_q}$ and n symmetric Cournot firms, the markup is $\mu = x/(x-1)$ with $x \equiv n\sigma_q$, and the producer price is $p = \mu c$, where c is marginal cost.

Free entry requires $(p-c)Q/n = F$, which after substitution reads

$$g(x) \;\equiv\; (x-1)^{\sigma_q-1}x^{-(\sigma_q+1)} \;=\; \frac{F}{A\sigma_q}\,c^{\sigma_q-1}(1+t_U)^{\sigma_q}.$$

The function g is single-peaked at $x = (\sigma_q+1)/2$; on the regular branch $x > (\sigma_q+1)/2$, equivalently $\mu < (\sigma_q+1)/(\sigma_q-1)$, it is decreasing, with $\mathrm{d}\ln g/\mathrm{d}\ln x = (\sigma_q-1)\mu - (\sigma_q+1) < 0$. Differentiating, $\mathrm{d}\ln x/\mathrm{d}\ln(1+t_U) = \sigma_q/[(\sigma_q-1)\mu-(\sigma_q+1)]$, and since $\mathrm{d}\ln\mu = -(\mu-1)\,\mathrm{d}\ln x$,

$$\frac{\mathrm{d}\ln\mu}{\mathrm{d}\ln(1+t_U)} = \frac{\sigma_q(\mu-1)}{(\sigma_q+1)-(\sigma_q-1)\mu}.$$

Adding the direct effect gives $\Phi = 1 + \mathrm{d}\ln\mu/\mathrm{d}\ln(1+t_U) = (\mu+1)/[2-(\mu-1)(\sigma_q-1)]$, after using $(\sigma_q+1)-(\sigma_q-1)\mu = 2-(\mu-1)(\sigma_q-1)$. For a change in marginal cost the right-hand side moves by $(\sigma_q-1)\,\mathrm{d}\ln c$ rather than $\sigma_q\,\mathrm{d}\ln(1+t_U)$, so $\mathrm{d}\ln\mu/\mathrm{d}\ln c = (\mu-1)(\sigma_q-1)/[2-(\mu-1)(\sigma_q-1)]$ and $\Phi_E = 1+\mathrm{d}\ln\mu/\mathrm{d}\ln c = [1-\tfrac{1}{2}(\mu-1)(\sigma_q-1)]^{-1}$, which exceeds one whenever $\sigma_q > 1$. ■

Proof of the pass-through identities (C.1)

At fixed n the equilibrium price is $p = \mu(n)\,c$ with μ constant under constant-elasticity demand. An emissions charge raises marginal cost, so $\mathrm{d}\ln p = \mathrm{d}\ln c$: proportional pass-through is one, which in levels means the buyer price rises by μ times the cost increase. The usage tax enters the buyer price as $q = (1+t_U)p$, so at fixed n $\mathrm{d}\ln q = \mathrm{d}\ln(1+t_U)$. Both identities hold for constant-elasticity demand and any conduct; with free entry both instruments also move n, as Proposition 6 shows. ■

Proof of Lemma 3

Let the reform induce $\mathrm{d}\ln w_j = \mathrm{d}\ln\bar{w} + \varepsilon_j$, with the relative component normalized by $\sum_j w_j L_j\,\varepsilon_j = 0$. Evaluate welfare at the optimal (τ, T): by the envelope theorem the induced re-optimization of the tax has no first-order effect, so we may compute the direct effects of the wage change holding (τ, T) fixed. Measured in units of public funds, welfare moves through two channels. By Roy's identity the private valuation of the wage change is $(1-\tau)\sum_j g_j\,w_j L_j\,\mathrm{d}\ln w_j$, where g_j is the social marginal welfare weight normalized so that its population mean is one. The induced change in revenue is $\tau\sum_j w_j L_j(1+\varepsilon_L)\,\mathrm{d}\ln w_j$, since a wage change of $\mathrm{d}\ln w_j$ raises earnings both mechanically and through the labor-

supply response ε_L that weak separability isolates; this revenue is returned lump sum.

The optimal linear tax satisfies the first-order condition

$$\sum_j g_j\, w_j L_j \;=\; \sum_j w_j L_j \Big(1 - \tfrac{\tau}{1-\tau}\,\varepsilon_L\Big),$$

i.e. the income-weighted average welfare weight equals $1 - \frac{\tau}{1-\tau}\varepsilon_L$. Apply this to the common component $\mathrm{d}\ln w_j = \mathrm{d}\ln \bar{w}$: its total contribution is

$$\mathrm{d}\ln \bar{w}\,\Big[(1-\tau)\sum_j g_j\, w_j L_j + \tau(1+\varepsilon_L)\sum_j w_j L_j\Big] = \mathrm{d}\ln \bar{w}\,\sum_j w_j L_j,$$

where the first-order condition collapses the bracket to $\sum_j w_j L_j$. This is precisely the marginal value of the change in aggregate labor income, the same object the resource accounting of Lemma 1 already books through $\mathrm{d}\ln Y$; it is an efficiency term, not a distributional one, and to avoid double counting it does not re-enter the distributional shadow value Ω. What remains for Ω is the relative component, $(1-\tau)\sum_j \hat{g}_j\, w_j L_j\, \varepsilon_j$, in which the common part has vanished because $\sum_j \hat{g}_j\, w_j L_j = 0$ by the definition of the income-weighted excess weights $\hat{g}_j \equiv g_j - \bar{g}$.

If the reform also shifts a distributional moment orthogonal to a proportional shift, a linear tax cannot neutralize it; writing its projection on the rank statistic as $\mathrm{Cov}_j(\hat{g}_j, \mathrm{rank}_j)$, a residual of order $(\mathrm{d}\ln \bar{w})\,\mathrm{Cov}_j(\hat{g}_j, \mathrm{rank}_j)$ remains, vanishing iff the covariance is zero. ■

Proof of Lemma D.2

Cost minimization gives $L_R \propto X_A (c_A/w_R)^{\sigma}$ within the automatable bundle and $X_A/X_N \propto (c_A/w_N)^{-\eta}$ across bundles, with $L_N = X_N$. Hence $L_R/L_N \propto c_A^{\sigma-\eta}\, w_N^{\eta}\, w_R^{-\sigma}$. Normalizing $\mathrm{d}\ln w_N = 0$, writing $x = \mathrm{d}\ln(w_R/w_N)$ and using $\mathrm{d}\ln c_A = (1-s)x + s\,\mathrm{d}\ln q$,

$$\mathrm{d}\ln(L_R/L_N) \;=\; (\sigma-\eta)s\,\mathrm{d}\ln q - \big[\eta + s(\sigma-\eta)\big]x.$$

Relative supply responds as $(\varepsilon_L + \mathcal{E})x$. Equating, and dropping the term of order s in the coefficient on x, which is the same wage feedback on q neglected in the text, gives $x = s(\sigma-\eta)\mathrm{d}\ln q/(\eta+\varepsilon_L+\mathcal{E})$. The remainder of the proof of Proposition 2 is unchanged: the factor s cancels against $p\,a = \lambda s Y$ as before. ■

Proof of Proposition B.2

The first part is the standard Pigouvian argument applied to K, together with the observation that the firm's choice of K is a decision about model capability, taken before and separately from the volume of inference sold, so the two are not proportional in general. For the second part, a threshold sets an implicit marginal tax of zero below $\bar{K}$ and infinity at $\bar{K}$. With heterogeneous providers, the marginal value of the last unit of compute differs across firms at the threshold, so the allocation of aggregate training compute across firms is not cost-minimizing for the risk attained. A uniform levy equalizes marginal value across firms at R', which is the standard equimarginal principle. ■

Proof of Proposition B.3

Write $H(T) \equiv \sum_{s=1}^{T} \rho^s$, which is positive and increasing in T. The efficiency first order condition (B.2) is $\varphi'(\xi) = -H(T)(p_e + t)e'(\xi)Q$. Define $G(\xi; a) \equiv \varphi'(\xi) + a\, e'(\xi)$ with $a \equiv H(T)(p_e + t)Q > 0$; the optimum solves $G = 0$. Since φ and e are strictly convex with $\varphi' > 0$, $\varphi'' > 0$, $e' < 0$, $e'' > 0$, we have $\partial G/\partial \xi = \varphi'' + a\, e'' > 0$, so the optimal ξ is unique, and $\partial G/\partial a = e' < 0$. By the implicit function theorem $d\xi/da = -(\partial G/\partial a)/(\partial G/\partial \xi) = -e'/(\varphi'' + ae'') > 0$. Thus ξ is increasing in a.

(i) The emissions charge sets $t = d > 0$, giving $a^E = H(T)(p_e + d)Q$, while the usage tax leaves the perceived energy price at p_e, giving $a^U = H(T)p_eQ < a^E$. Since ξ increases in a, $\xi^E(T) > \xi^U(T)$ for every $T \geq 1$.

(ii) a is increasing in T through $H(T)$, so both efficiencies rise with the horizon. For the gap, write $\xi^E - \xi^U = \int_{a^U}^{a^E} \xi'(a)\, da$ with $\xi'(a) = -e'/(\varphi'' + ae'') > 0$ and $a^E - a^U = H(T)dQ$, so the interval of integration widens in T while its left endpoint $a^U = H(T)p_eQ$ also moves right. Two forces therefore act in opposite directions, and the gap is increasing in T if and only if the widening dominates the decay of ξ'. A sufficient condition is that $a\,\xi'(a)$ be nondecreasing, i.e. that the elasticity of ξ with respect to the effective energy price not fall as that price rises; this holds, for instance, in the isoelastic case $\varphi(\xi) = \varphi_0\xi^{1+\nu}$ with $e(\xi) = e_0\xi^{-\epsilon}$, for which $\xi \propto a^{1/(1+\nu+\epsilon)}$ and the gap scales as $H(T)^{1/(1+\nu+\epsilon)}\big[(p_e + d)^{1/(1+\nu+\epsilon)} - p_e^{1/(1+\nu+\epsilon)}\big]Q^{1/(1+\nu+\epsilon)}$, manifestly increasing in T. We adopt this condition; absent it, part (i) and part (iii) below are unaffected, and only the monotonicity of the efficiency gap itself is at issue.

(iii) The per-period social cost of provision at efficiency ξ is $(p_e + d)e(\xi)Q$, and the one-time investment cost is $\varphi(\xi)$. Total social cost over the horizon is $C(\xi; T) = \varphi(\xi) + H(T)(p_e + d)e(\xi)Q$, minimized at $\xi^E(T)$. Evaluated at the usage-tax choice $\xi^U(T)$, which

is not the minimizer, the excess $\Delta(T) = C(\xi^U(T); T) - C(\xi^E(T); T) \geq 0$. Differentiating and applying the envelope theorem at ξ^E, $d\Delta/dT = (p_e + d)Q\left[e(\xi^U) - e(\xi^E)\right] H'(T) > 0$ since $e(\xi^U) > e(\xi^E)$ by (i) and $H'(T) > 0$. Convexity does not follow in general, since the bracket $e(\xi^U(T)) - e(\xi^E(T))$ can shrink as both efficiencies rise, and we do not claim it. ■

The argument behind Remark 3

Under the cash flow tax of Proposition 4, the provider's optimum is unchanged and after-tax profit is $(1 - z)\Pi$, so incumbent profit falls by $z\Pi$, exactly the revenue raised. With entry blocked, a usage tax leaves the producer price unchanged under constant-elasticity demand and lowers profit through the volume lost, so incumbents prefer neither. With free entry in every period, profit is zero under both. With incumbents' costs sunk and entry free only for the next generation, a usage tax lowers the number of next-generation firms at the rate of Proposition 6 and so raises next-generation markups; incumbents who expect to survive gain that margin on retained sales. Whether the gain exceeds the current-period burden depends on the discount factor and on how incumbency carries over, which we do not model, so the statement is a possibility rather than a prediction. ■

B Extensions

This appendix collects extensions referenced in Section 5. None changes the sign of any term in (8), and all move the optimal usage tax in the same direction, which is down.

B.1 A caveat on the form of the tax

The argument so far concerns the tax *base*: energy versus the value of AI output. There is a separate question about the tax *form*, namely whether a levy is specific (a fixed amount per kilowatt-hour or per token) or ad valorem (a percentage of price), and market power bears on it in a direction that partly cuts against our recommendation. Since we have leaned on market power to argue against the usage tax, we owe the other half of the ledger.

The classical result is due to Suits and Musgrave (1953) and, for oligopoly, Delipalla and Keen (1992): for equal revenue, an ad valorem tax on an imperfectly competitive

market yields a lower consumer price than a specific tax, and is therefore welfare superior. The comparison is in fact one of the oldest questions in public finance, going back to Cournot (1838) and surveyed by Keen (1998), who traces the line from Cournot's monopoly analysis through Wicksell to the modern competitive statement of Suits and Musgrave (1953). The mechanism is that an ad valorem tax scales with price, so the treasury absorbs a fixed share of any price increase the firm engineers, which blunts the firm's incentive to restrict output. A specific tax has no such property; the firm keeps the full price increase net of a constant per-unit charge. Under perfect competition the two forms are equivalent, but the AI sector is a case where they are not.

This matters because our preferred instrument, a per-kilowatt-hour charge at the data center, is specific, while the usage tax we argue against is ad valorem. On the market-power margin alone, then, the instrument we favor is the less well behaved of the two. The tension is real, but it does not overturn the ranking, for four reasons. First, the two considerations operate on different objects: the base question is about *targeting* the externality, which is worth the order of magnitude documented in Section 6, while the form question is a second-order refinement of pass-through within a given base. A large gain on targeting dominates a smaller loss on form. Second, the Suits–Musgrave advantage of the ad valorem form is largest when it applies to the marked-up good, and the data center's electricity is not the marked-up good; the markup sits downstream, on the token. A charge on the upstream input inherits little of the downstream market power, so the form penalty is muted where we actually propose to levy. Third, the efficiency-margin and location results of Proposition 3 and Corollary 3 have no counterpart for the ad valorem form: an ad valorem tax on tokens still cannot move the carbon intensity of a kilowatt-hour, whatever its pass-through properties. Fourth, and although we do not model electricity use in other sectors, a charge on data center power would in practice be a step toward (or an extension of) a charge on industrial electricity generally. Applied broadly it corrects the same externality without introducing a wedge between AI and other users of the same kilowatt-hour, which is precisely the inter-sectoral distortion an AI-specific instrument risks creating. Section 6.6 takes up how AI's electricity use interacts with existing rate structures, and the point there reinforces this one: the reforms that matter most are changes to how large loads are charged in general, not a levy carved out for AI.

The clean statement is therefore as follows. If the choice is between a specific and an ad valorem tax *on the same base*, market power favors the ad valorem form (Keen, 1998). But the first-order choice is the base, not the form, and there the externality argument points decisively at energy. Where the two principles can both be honored, they should

be: a levy on data center energy can itself be structured as a percentage of the wholesale power price rather than as a flat per-kilowatt-hour charge, capturing the Suits–Musgrave benefit on the small amount of market power present in generation while keeping the base correct. Indeed Myles (1996) shows that a suitable *combination* of a specific and an ad valorem instrument can reproduce the Ramsey price and undo the imperfect-competition welfare loss outright, which is the design-space version of the same point. We flag this as a design refinement rather than build it into the model, since it does not affect any of the magnitudes.

B.2 Training is not sunk once: recurring investment, the rent base, and the regime

A sunk-stock reading of frontier provision is convenient but not credible. It requires that no further training ever be necessary, whereas frontier providers retrain continuously: each model generation requires a new run, post-training and evaluation are ongoing, and clusters depreciate both physically and through obsolescence as accelerator generations turn over. The right frame is ex ante and stationary rather than ex post and one-shot.

Let providers spend I per period on training and capacity to remain at the frontier, and write $\phi \equiv I/(p\,a)$ for recurring investment as a share of revenue. Under a cash flow tax with immediate expensing, I is deducted as incurred, so the base is the operating surplus net of investment:

$$\text{cash flow base} \;=\; \Big(1-\tfrac{1}{\mu}\Big)p\,a \;-\; I \;=\; \Big[\Big(1-\tfrac{1}{\mu}\Big)-\phi\Big]\,p\,a. \tag{B.1}$$

Two consequences follow. First, the rent tax raises less than the sunk-stock reading suggests. At the central $\mu = 1.5$ the operating surplus is a third of revenue, so at $\phi = 0.3$ the taxable residual is about three percent of revenue and a fifty percent cash flow tax raises a few hundredths of a percent of GDP rather than the 0.25 percent of Table 6, which taxes the operating surplus as if training were sunk. Expensing still preserves neutrality: the marginal training run is never taxed.

Second, recurring investment indicates which regime applies; it does not scale the market-power term. If recurring investment absorbs most of the operating surplus, economic profit is close to zero, which is what free entry delivers. At the central markup, $\phi \in [0.2, 0.4]$ puts economic profit between about -7 and $+13$ percent of revenue. The evidence on recurring training therefore points to the free-entry regime at the horizon over which training recurs, and to the fixed-provider regime only within a model generation.

In neither regime does the fixed cost shrink the term: with entry blocked it is sunk and the whole margin is underprovision, and with free entry the margin pays for entry and the term reflects how a tax raises the markup (Section 5.6).

B.3 Mobile compute and the case for a destination base

The one argument that survives is the one least often made. Compute is mobile. A charge on data center emissions can be avoided by moving the load, and the emissions then occur abroad, often on a dirtier grid.

Proposition B.1 (Leakage and border adjustment)**.** *Let $\ell \in [0,1]$ be the leakage rate, the fraction of the reduction in domestic data center emissions that reappears abroad in response to a unilateral charge, and suppose the planner values global damages. Then:*

(i) the optimal origin-based charge is $t_E^\star = (1-\ell)\, d$, short of the full Pigouvian rate;

(ii) a destination-based levy on AI services consumed domestically but served from abroad, at the ad valorem rate D/p on those services with credit for charges paid at origin, prices their emissions and cannot be avoided by relocating the compute that serves the domestic market;

(iii) with that levy in place, relocation no longer avoids the charge for domestic consumption, so the leakage that motivates (i) is removed for that consumption and the origin charge can be set at the full rate d.

This is the strongest available defence of a tax on AI usage, and it is narrow. The tax is optimal not because AI is an input worth discouraging, not because it displaces labor, and not because its providers are rich. It is optimal because it is the only administratively feasible way to price a physical externality that has been shipped offshore. It should therefore be designed as a border adjustment: benchmarked to the embodied energy and grid intensity of the compute, creditable against carbon prices paid abroad, and set to zero for compute served from a jurisdiction with an equivalent charge. A usage tax without these features is not implementing Proposition B.1.

There is a working precedent, and it is instructive in both directions. The European Union's Carbon Border Adjustment Mechanism applies exactly this logic to goods: importers of cement, steel, aluminium, fertiliser, electricity and hydrogen must surrender certificates matching the embedded emissions of what they bring in, with credit for carbon prices already paid in the country of origin. The design principles we list above are the ones CBAM adopts, which is some evidence they are administrable rather than merely

coherent. But CBAM also shows what the theory costs in practice. Embedded emissions must be reported, verified, and defaulted when unverifiable; the definitional perimeter is contested; the mechanism took years to phase in and remains disputed by trading partners. Those difficulties arise for physical goods that cross a border on a ship with a bill of lading. An AI query has no such paper trail, and, as Section E.1 argues, attributing embodied energy to a service routed dynamically across data centers is materially harder than attributing it to a tonne of steel. We take the CBAM experience as evidence that a border adjustment on embodied emissions is possible, and as a warning that our version of it is the least implementable recommendation in this paper.

B.4 Training compute and risk

Inference and training are different activities with different externalities, and an instrument levied on tokens taxes almost exclusively the former. Suppose the largest training runs generate an external social risk $R(K)$, convex in the compute K used in a single run, whether from misuse, from concentration, or from failures of control. Let training compute be purchased at price p_K.

Proposition B.2 (Training levy)**.** *The corrective instrument is a Pigouvian levy on training compute at the rate $t_K^\star = R'(K)$, evaluated at the post-tax level of K. A usage tax on inference implements $t_K^\star$ only in the degenerate case in which training compute is used in fixed proportion to inference volume. Moreover, when risk depends on aggregate training compute, a uniform levy attains any expected risk a compute threshold $\bar{K}$ attains at no higher resource cost, and at strictly lower cost whenever providers' net marginal values of compute differ at the threshold allocation. Heterogeneity in the value of scale is necessary for this but not sufficient; when risk is firm-specific and convex in a firm's own compute, the levy must track each firm's marginal risk to retain the advantage.*

The compute thresholds already written into regulation are a crude version of this instrument. The model says that a price would dominate a threshold, for the usual reason that a threshold is a tax rate of zero followed by a tax rate of infinity, and that the base should be training compute rather than tokens served. Two caveats belong on the record.[B.1]

[B.1] Convexity of R is an assumption, not a finding; and if the marginal risk is not convex but has a threshold shape, the ranking can reverse, which is presumably the implicit theory behind the current rules.

There is a second and more prosaic reason to separate training from inference, and it belongs with the environmental argument rather than the risk argument. Training consumes electricity and emits carbon, but it produces *no tokens*. The externality calculation of Section 6 is computed per query and therefore covers inference only; the energy burned training the model is nowhere in the base of a token tax. Its implicit tax rate is not low but exactly zero. Estimates of the inference share of AI compute energy vary widely, historical figures ran from about a third to sixty percent, and recent industry estimates put inference at eighty to ninety percent of a deployed model's lifetime energy, so somewhere between one tenth and two fifths of the sector's electricity is invisible to a levy on tokens. A charge at the meter reaches it automatically, because a kilowatt-hour drawn to train a model and a kilowatt-hour drawn to serve a query are billed identically. This is a distinct point from the mis-scaling of Table 1: there the token tax applies the wrong rate to the harm, here it applies no rate at all to a large part of it. Because training is part of the fixed cost of entry, a meter charge on it also raises F and so reduces entry, a channel the instrument comparison does not track; this does not weaken the case for pricing training emissions, and conversely a token tax's zero rate on training electricity does not mean it leaves training untouched, since it too affects entry. Given that training is recurring rather than sunk, as Section B.2 argues, that blind spot is permanent rather than transitional.

B.5 Dynamic efficiency: why the gap grows

Proposition 3 compares the two instruments at a fixed efficiency technology. But the efficiency of AI provision is not fixed; it is the outcome of investment, and part of the striking fall in energy per token over recent years plausibly reflects the cost of energy, although cheaper chips, better serving and distillation lower compute costs as well, so the historical decline does not by itself measure the investment an electricity charge would induce. Once efficiency is a durable investment rather than a static choice, the two instruments diverge by more, and the divergence compounds over time. This is the one place where the static formula understates its own conclusion.

Nothing here requires a long horizon. The essential structure is two periods, an ex ante date at which the durable efficiency investment is chosen and an ex post date at which output is served.[B.2] We write the model with T periods because the horizon is itself the object of interest: it is what makes the instrument gap grow rather than merely exist.

[B.2]The reader who prefers that reading can set $T = 2$ throughout and lose only the comparative static in part (ii) below.

Extend the model to T periods, then. In period one, providers make a durable investment in efficiency ξ at cost $\varphi(\xi)$, which lowers energy per unit $e(\xi)$ in every subsequent period. Energy is purchased period by period at price p_e, and the social cost of a kilowatt-hour is $p_e + d$. A provider facing an instrument that raises its perceived energy price to $p_e + t$ chooses

$$\varphi'(\xi) \;=\; -\left(\sum_{s=1}^{T} \rho^s\right)(p_e + t)\, e'(\xi)\, Q, \tag{B.2}$$

where ρ is the discount factor and Q is per-period output. The bracketed horizon term multiplies the energy price, so a longer horizon raises the investment in efficiency for any given t.

Proposition B.3 (The dynamic instrument gap)**.** *Let $\xi^E(T)$ be the efficiency chosen under an emissions charge $t_E = d$ and $\xi^U(T)$ that chosen under a revenue-equivalent usage tax, which leaves the perceived energy price at p_e, comparing the two at the same scale of AI use. Then:*

(i) $\xi^E(T) > \xi^U(T)$ for all $T \geq 1$: the emissions charge always induces more efficiency investment, because only it raises the perceived price of energy.

(ii) Both are increasing in the horizon. The gap $\xi^E(T) - \xi^U(T)$ is increasing in T if the elasticity of efficiency with respect to the effective energy price does not fall as that price rises, as in the isoelastic case; without that condition it need not be.

(iii) The cumulative welfare gap between the instruments, $\Delta(T)$, is increasing in T: a usage tax matched to the emissions charge on the scale margin falls further behind on the efficiency margin as the horizon lengthens.

The proof is in Appendix A; the mechanism is entirely in (B.2). A usage tax is a wedge on output, so it never enters the energy price and never appears in the efficiency condition, at any horizon. An emissions charge enters (B.2) directly and scales up with the life of the asset. The efficiency margin is where the observed order-of-magnitude gains in energy per token came from, so an instrument that is neutral toward it is not neutral toward the thing that has actually reduced the footprint of AI.

The dynamic result sharpens the policy reading in a way worth stating plainly. A common defense of the token tax is that it is simple and the emissions data are noisy, so the token tax is a reasonable approximation to be tightened later. Proposition B.3 says the approximation gets worse, not better, with time, because it forgoes an investment response that accumulates. This argues for getting the base right early, even crudely, because the base is what steers durable investment. The proposition compares instruments maintained

over an asset's life rather than policy paths that switch later, so it supports concern about persistent early underinvestment rather than a full ranking of sequencing strategies.

B.6 The market-power term under free entry

Proof of Proposition 7. The argument of the proof of Proposition 2 applies with provider profit weighted by g_S rather than one: the profit term becomes $(1-f)g_S(p-m)\,da$, and the first-order condition gives the stated rate. The shareholder weight enters because, with the number of providers fixed, the margin on a marginal unit accrues to the owners of the providers. ■

Proof of Proposition 8. With free entry, profit is zero in every equilibrium, so domestic welfare is $W = CS(q) + t_U\, p\, Q - (\Omega + D - \psi)\, Q$ and ownership does not enter. Write $u = \ln(1+t_U)$. By Proposition 6, $\mathrm{d}\ln q/du = \Phi$, so $dQ/du = -\sigma_q \Phi Q$ and $\mathrm{d}\ln p/du = \Phi - 1$. Then

$$\frac{dW}{du} \;=\; pQ\Big[(1+t_U)(1-\Phi)+t_U(\Phi-1-\sigma_q\Phi)+\sigma_q\Phi\,\frac{\Omega+D-\psi}{p}\Big] \;=\; pQ\Big[(1-\Phi)-t_U\sigma_q\Phi+\sigma_q\Phi\,\frac{\Omega+D-\psi}{p}\Big],$$

and setting this to zero at the values of Φ and p prevailing at $t_U = 0$ gives the stated expression. Because Φ rises as the tax shrinks the market, the exact optimum differs from the first-order expression at high markups. Table B.1 reports it, found by maximizing W over the solved free-entry equilibrium. The implied share of the textbook correction that entry absorbs, $\chi \equiv 1 - [(\Omega + D)/p - t_U^\star]/(1 - 1/\mu)$, applies to the ad valorem usage tax; it differs from the share for an output subsidy, because the tax also changes the producer revenue that determines entry. ■

Table B.1: Free entry: exact optimum at the central calibration, $(\Omega + D)/p = 6.44\%$

μ	firms n	Φ	market-power term	textbook $1 - 1/\mu$	implied χ	$t_U^\star$
1.2	3.05	1.22	8.8%	16.7%	0.47	−2.4%
1.5	1.53	1.65	16.6%	33.3%	0.50	−10.2%
2.0	1.02	2.90	22.1%	50.0%	0.56	−15.6%

B.7 Illustrative calculations

Revenue, distortion and output. Table 6 runs each instrument through the model at the central calibration. The rates correspond to instruments that have been proposed rather than to optima: a ten percent usage tax as the round number that appears in

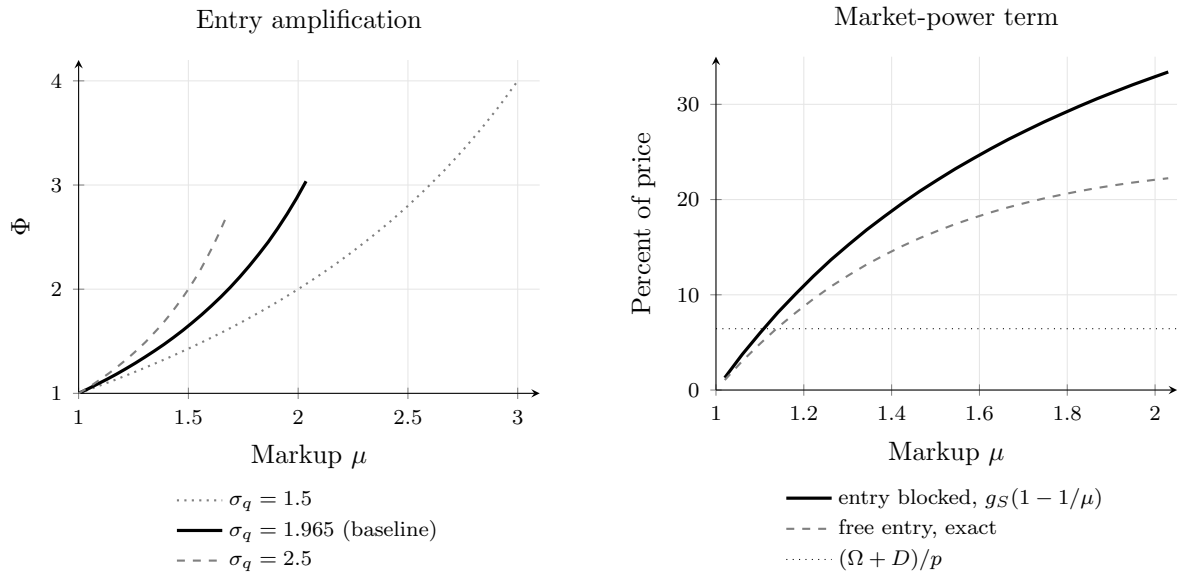


Figure B.1: Left: the buyer-price response to a usage tax under free entry, $\Phi = (\mu+1)/[2-(\mu-1)(\sigma_q-1)]$, on the regular branch and within the Lerner bound $\mu \le \sigma_q/(\sigma_q-1)$. Right: the market-power term at the baseline σ_q, with entry blocked and under free entry, against the sum of the predistribution and Pigouvian terms. The domestic optimal usage tax is negative wherever the relevant market-power curve lies above the dotted line, which happens at any markup above about ten percent.

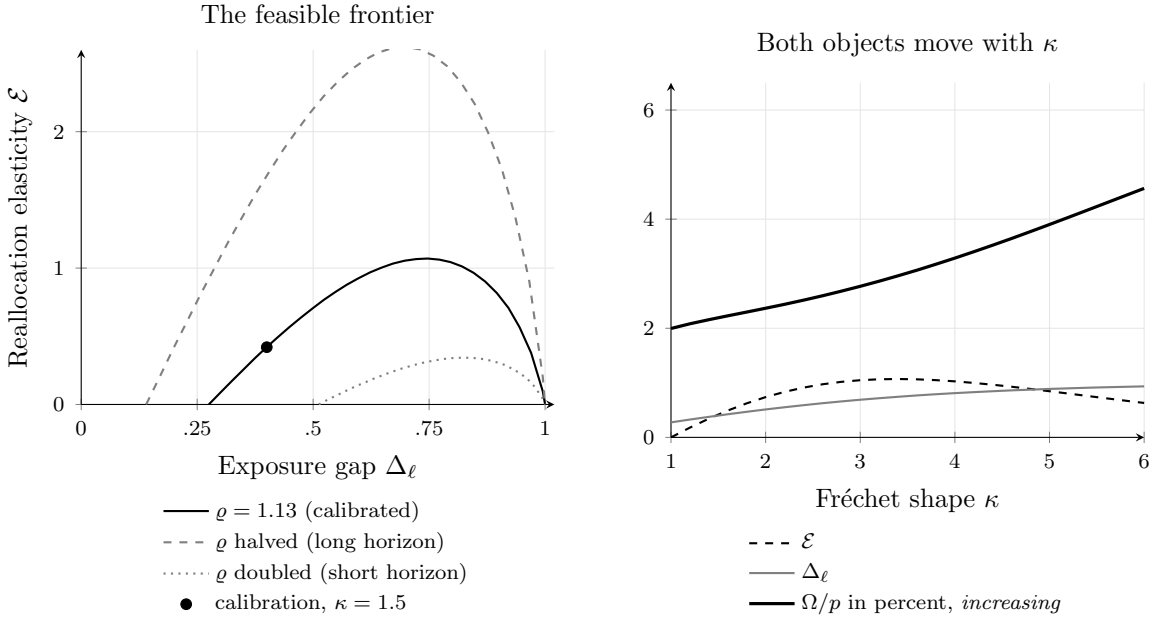


Figure B.2: Proposition 5 drawn, in the symmetric case and with reallocation in efficiency units. Left: for each comparative-advantage gap ϱ, the pairs $(\Delta_\ell, \mathcal{E})$ the Roy model can generate, with $\mathcal{E} = (4\,\mathrm{artanh}\,\Delta_\ell/\varrho - 1)(1-\Delta_\ell^2)$ for $\kappa > 1$. Each frontier starts from zero at $\Delta_\ell = \tanh(\varrho/4)$, where effective reallocation vanishes, and returns to zero at $\Delta_\ell = 1$. The frontier constrains the pair for a given ϱ; a larger ϱ can deliver sharper identification and slower reallocation together, so the case against the fixed-assignment corner is empirical, resting on the measured gap, rather than a theoretical exclusion. Comparative advantage is sharpest early and erodes as cohorts adapt, so the larger ϱ describes the short horizon. Right: raising κ at the calibrated $\varrho = 1.13$ sharpens sorting faster than it speeds reallocation, so the predistribution term rises monotonically toward its fixed-assignment value.

public discussion, three percent as the modal rate among existing digital services taxes, a fifty percent cash flow tax on rents, and a charge of $d = \$0.05$ per kilowatt-hour on data center load. Each entry is the full change in domestic welfare, counting the pre-existing markup, the externality and the wage effects alongside the tax. Three things stand out. The rent tax raises about twice what a ten percent usage tax raises with entry blocked, at no first-order cost, but has no base under free entry. The usage tax costs 0.42 dollars of welfare per dollar raised for a domestic sector with entry blocked, because it deepens an underprovision that already exists. And the emissions charge, scored on the scale margin alone, also loses welfare under constant-elasticity demand, because the sector passes it through with amplification; its advantages lie on the efficiency and training margins the scoring holds fixed.

B.8 Model primitives held over from Section 3

The markup requires interpretation. A firm that charges above short-run marginal cost is not thereby earning rent: with a large fixed cost F, free entry drives profit to zero and the entire margin recovers F. Section 5.6 shows that this does not remove the distortion on the margin, and that what the market-power term measures depends on whether entry is blocked or free.

The second feature is the hinge of the targeting argument. There are two distinct margins by which the external cost of a unit of AI can fall: the *scale* margin, fewer units bought, and the *efficiency* margin, fewer joules per unit. A wedge on output reaches only the first. A charge on energy reaches both, because it enters the provider's marginal cost through $p_e e(\xi)$ and so changes the efficiency ξ that minimizes it. This is why Proposition 3 is not merely the standard targeting principle restated: the instruments differ not only in precision but in the set of margins they can touch at all.

Two objects follow from the Fréchet assumption and carry the whole distributional analysis. The share of group i working in routine tasks and the group's expected wage index are

$$\ell_{iR} \;=\; \frac{(A_{iR}w_R)^{\kappa}}{(A_{iR}w_R)^{\kappa}+(A_{iN}w_N)^{\kappa}}, \qquad W_i \;=\; \Gamma\big(1-\tfrac{1}{\kappa}\big)\left[(A_{iR}w_R)^{\kappa}+(A_{iN}w_N)^{\kappa}\right]^{1/\kappa}, \tag{B.3}$$

and differentiating W_i gives the statistic that replaces "the routine wage":

$$\mathrm{d}\ln W_i \;=\; \ell_{iR}\,\mathrm{d}\ln w_R+(1-\ell_{iR})\,\mathrm{d}\ln w_N, \qquad \text{hence} \qquad \mathrm{d}\ln\frac{W_1}{W_2} \;=\; \Delta_\ell\,\mathrm{d}\ln\frac{w_R}{w_N}, \qquad \Delta_\ell \;\equiv\; \ell_{1R}-\ell_{2R}. \tag{B.4}$$

Equation (B.4) is an envelope result: the marginal switcher is indifferent between tasks, so reallocation has no first-order welfare effect and the entire burden of a relative task-wage movement falls on inframarginal workers, in proportion to the share of the group that is exposed. We call $\Delta_\ell \in [-1, 1]$ the *exposure gap*. With $\varrho > 0$ group 1 has comparative advantage in exposed tasks and $\Delta_\ell > 0$; when group 1 is instead defined as the group the planner weights more heavily, as in the empirical work of Section 6, a negative measured gap corresponds to $\varrho < 0$, and the predistribution term changes sign with it while $\hat{g}_1$ stays positive. It is the fraction of a relative task-wage movement that survives as a relative movement between the two groups the planner actually cares about, and it is in principle measurable.

B.9 The shareholder welfare weight

The calibration is disciplined by two observable positions. Routine occupations sit near the 36th percentile of the earnings distribution, $z_R \approx -0.35$, and the dispersion of log earnings is $s \approx 0.8$; these give $\eta \approx 0.32$. Corporate equity is held overwhelmingly at the top—the top one percent hold roughly half of it—so the equity-weighted dollar sits near $z_S \approx 2.3$, which gives

$$g_S \;\approx\; 0.65, \qquad \text{range } 0.45 \text{ to } 0.85$$

across $s \in [0.7, 1.0]$, $z_R \in [-0.5, -0.25]$ and $z_S \in [2.0, 3.0]$.

The calibration is a constraint rather than a licence. A planner who values a routine worker only thirty percent above the average is not strongly redistributive, and the same modest curvature that keeps the predistribution term small also keeps shareholders from being heavily discounted: $\hat{g}_1 = 0.3$ and $g_S = 0.2$ are not jointly available under any isoelastic welfare function. With entry blocked, the market-power term is therefore about two thirds of what (8) reports with profit weighted at one.

It also answers the objection that the deadweight loss from market power might simply be smaller than that from a tax, so the comparison should run the other way. That compares levels; the formula is a marginal condition. What matters is not the size of the Harberger triangle under the monopoly wedge but that a wedge already exists where the tax is levied: a tax imposed on top of a pre-existing wedge has a *first-order* welfare cost rather than the second-order cost it would have in an undistorted market. That is why the term enters linearly in m rather than quadratically, and the sense in which the two distortions compound rather than compete.

B.10 The mobility–identification frontier

The intuition. Think of the Fréchet shape parameter as answering a single question: how different are workers from one another *within* a group? When dispersion is wide, everyone is a coin flip between tasks, the two groups look alike, and there is no identifiable loser for the planner to protect—but there is a great deal of switching, because half the workforce sits near the margin. When dispersion is narrow, comparative advantage is nearly deterministic: group 1 works routine tasks and group 2 does not, so the loser is perfectly identified—and nobody switches, because nobody is near the margin. The same parameter that makes workers mobile is the parameter that makes them indistinguishable. It cannot be turned up to damp the wage effect while the welfare weight on "routine workers" is held fixed, because turning it up is what creates a routine worker in the first place.

Part (iv) is where the economics is. The right way to discipline the term is not to argue about κ but to measure the two objects the formula actually contains. The exposure gap Δ_ℓ is the difference between two employment shares, the fraction of the disadvantaged group working in AI-exposed tasks, less the fraction of the advantaged group doing so, and it is estimable from any occupational exposure measure crossed with any ex ante grouping the planner might have in mind. The reallocation elasticity $\mathcal{E}$ is then pinned down by Lemma 2 given a switching estimate. The literature's convention $\Delta_\ell = 1$ is not a conservative simplification. It is the corner of the parameter space that maximizes the term, and it is empirically false.[B.3] Section 6 measures it, carries $\Delta_\ell \approx 0.4$—the value on the exposure measure most favourable to the robot-tax case, with every direct AI measure giving a negative number—and finds the term falls to about a fifth of the fixed-assignment number. The reduction comes chiefly from the exposure gap rather than from mobility.

Proposition 5 says the pair $(\Delta_\ell, \mathcal{E})$ is not free. The frontier is drawn in the left panel of Figure B.2; it is the object a calibration should be disciplined against, and it makes clear which combinations of beliefs are actually available. Three readings follow.

In the symmetric case, eliminating κ between $\Delta_\ell = \tanh(\kappa\varrho/4)$ and $\mathcal{E} = (\kappa-1)(1-\Delta_\ell^2)$ gives

$$\mathcal{E} \;=\; \Big(\frac{4}{\varrho}\,\mathrm{artanh}(\Delta_\ell) - 1\Big)\big(1-\Delta_\ell^2\big),$$

which is zero at $\Delta_\ell = \tanh(\varrho/4)$, where effective reallocation vanishes, rises, and returns to zero at $\Delta_\ell = 1$.

[B.3] On the education and job-preparation groupings of Table 2, exposure to AI is far from segregated; earnings and other groupings would require individual-level microdata, as Appendix D.9 discusses.

First, for a given comparative-advantage gap the feasible set excludes the configuration on which the robot-tax case implicitly rests, though a larger gap can relax that exclusion, so the case against it is ultimately empirical. That case wants a large exposure gap together with slow reallocation, so that the wage effect is not competed away.[B.4] The frontier says these are compatible only in the high-Δ_ℓ corner, and in that corner the model is the fixed-assignment model, whose empirical premise is that AI exposure is perfectly segregated across ex ante groups. It is not.

Second, the horizon enters through ϱ rather than through κ. Over a quarter, comparative advantage is whatever accumulated human capital made it, and ϱ is large. Over a decade, cohorts choose skills knowing what AI can already do, which compresses ϱ toward zero and takes Δ_ℓ with it. This is a cleaner statement of the transitional argument than the one in terms of a rising mobility parameter: what decays is not the friction but the identifiability of the losing group, and the predistribution term decays with it. The right sunset is the horizon over which ϱ compresses, which depends on cohort turnover and on how fast the exposure frontier itself moves.

Third, two caveats cut in opposite directions and neither is resolved here. Reallocation is not free even when it happens: switching costs, foregone tenure, and human capital specific to the abandoned task are real resource losses that a relative-wage statistic does not capture, so a positive $\mathcal{E}$ understates the welfare loss from displacement even as it damps the measured wage effect. Against that, the exposure gap is measured on *ex ante* groups, and if the planner's actual normative concern is with individuals rather than groups, with whoever happens to hold strong comparative advantage in a routine task, then the two-group structure still overstates how cleanly the burden falls on an identifiable bloc, which is the limitation Thümmel (2023) identifies. What survives both is that the redistributive case for taxing AI usage is weaker, more transitional, and more dependent on an unmeasured sorting pattern than the two-type model implies. We carry Δ_ℓ explicitly from here on, report it as a range, and never rely on the $\Delta_\ell = 1$ corner alone.

B.11 The zeros of Proposition 4 in practice

Neutrality is a property of an idealized cash flow base and no enacted statute has had all of its features. Four channels put a positive wedge back in, and each has a known drafting remedy.

- *Imperfect loss offset.* If losses are carried forward without interest, or expire, the

[B.4] A large gap means a sharply identified group of losers, so that $\hat{g}_1$ has something to attach to.

state takes a full share of the upside and less than a full share of the downside. This is a tax on risk-taking and it falls hardest on exactly the frontier investments whose returns are most dispersed. Remedy: carry-forward with interest at the government's borrowing rate, or refundability.

- *An incomplete definition of the base.* Anything not expensed[B.5] leaves part of the normal return inside the base, and the tax then reaches marginal as well as infra-marginal returns.
- *Entity mobility.* A cash flow tax on a legally resident entity is neutral with respect to real decisions but not with respect to where the entity is resident. The base is stickier than a residence-based profits tax only to the extent that the underlying compute is physically located, which is one more argument for the meter.
- *Time inconsistency.* A tax that is neutral because the capital is sunk is one a rational investor expects to be repeated, and the anticipation raises the required return on new investment. This is the sharpest objection to the sunk-stock reading of the base, and it is why §B.2 prefers the stationary reading.

B.12 Diffusion spillovers

If adoption generates learning spillovers, so that the social return to AI use exceeds the private return by ψ per unit, the corrective term becomes $(D - \psi)/p$. We do not attempt to sign ψ: knowledge spillovers push it up, business stealing pushes it down, and the empirical literature is young. What can be said is the direction of its effect on policy. Any positive ψ lowers the optimal usage tax, and it lowers it against a Pigouvian term that Section 6 puts at around three percent, so a spillover of that size would cancel it. It does not take a large spillover to cancel the environmental case for a token tax.

Collecting the extensions, the general expression is

B.13 Recurring investment: derivations

The base (B.1) follows from immediate expensing: the operating surplus is $(1 - 1/\mu)\, p\, a$, recurring investment $I = \phi\, p\, a$ is deducted as incurred, and the residual is economic profit. Its sign is what separates the two regimes of Section 5.6: free entry drives it to zero, while blocked entry leaves it positive. At $\mu = 1.5$ and an AI cost share of $\theta = 1.5$ percent of GDP, the operating surplus is 0.5 percent of GDP; a fifty percent tax on it raises 0.25

[B.5] Capitalized software, acquired intangibles, and capitalized research under the current U.S. treatment.

percent, and a fifty percent tax on the stationary base raises $0.5\,(1/3-\phi)\times 1.5$ percent of GDP, about 0.03 percent at $\phi = 0.3$ and nothing at $\phi \geq 1/3$.

C Derivations and interpretation

This appendix develops three results stated compactly in the body.

C.1 The negative wedge and the role of ownership

Does the formula call for a subsidy? Under constant-elasticity demand it does, and we should say so plainly. Set $\hat{g}_1 = 0$ or $\sigma = 1$ and $D = \psi = 0$: with entry blocked, (9) returns $t_U^\star = -(1-f)\,g_S(1-1/\mu) < 0$ whenever $f < 1$, and under free entry it returns approximately $-(\Phi-1)/(\sigma_q\Phi) < 0$ whatever the ownership. This is the oldest result in the economics of monopoly: a firm pricing above marginal cost produces too little, and a subsidy that closes the gap restores the competitive quantity (Cournot, 1838).

A theoretical case for a subsidy is not a policy proposal, and the reasons for not recommending one are practical. Implementing it requires knowing the markup over marginal cost and the pass-through of taxes into prices, neither of which has been measured for AI services, and Section 6.5 shows the sign itself turns on the second. It would hand public money to a concentrated sector, which invites the political economy of Appendix E.2. And the distortion it addresses has better instruments: entry, interoperability and procurement policy act on the markup directly, whereas a subsidy pays for a symptom.

Why ownership matters, and when it stops mattering. With entry blocked, the producer price does not respond to a usage tax, so a marginal tax change falls entirely on buyers whose loss on inframarginal units is the revenue collected; what remains is the margin on units no longer sold, which is profit. Profit accruing abroad is not domestic welfare, so the importer's optimal rate exceeds the domestic one by the shareholder-weighted Lerner index. Under free entry profit is zero, the gain from cheaper AI accrues to buyers, and ownership drops out: an importer then wants the same rate as a country that owns its providers. The ordering by ownership is the one result that survives the change in demand curvature of Section 6.5.

C.2 Interpreting the markup: fixed-cost recovery and timing

A firm that charges above marginal cost is not thereby earning rent: with a large fixed cost F, free entry drives profit to zero and the entire margin recovers F. That does not remove the distortion on the margin. Whether the margin is rent or fixed-cost recovery, a buyer who values a unit above its marginal cost but below its price does not buy it. What differs between the two regimes is what a tax does. With the number of providers fixed, F is sunk and a usage tax deepens the underprovision one for one. With free entry, F is what the margin pays for, and a usage tax additionally drives out entrants and raises the markup, which is the amplification of Proposition 6.[C.1]

C.3 Pass-through, entry, and the evidence on grid cost shifting

This appendix collects material referenced in Sections 2 and 3: the fuller statement of the pass-through discipline, the caveats to the entry-amplification result, and the detail of the ratepayer evidence.

Pass-through. Because the same imperfectly competitive sector must deliver both the targeting result and the entry result, we fix the pass-through of taxes into prices once, from (4), and use it consistently. With constant-elasticity demand and the conduct in (4), the pass-through of a cost shock (such as the emissions charge, which raises m) and of the ad valorem usage tax are, respectively,

$$\rho_E \;\equiv\; \left.\frac{\mathrm{d}\ln p}{\mathrm{d}\ln m}\right|_{n \text{ fixed}} \;=\; 1, \qquad \rho_U \;\equiv\; \left.\frac{\mathrm{d}\ln q}{\mathrm{d}\ln(1+t_U)}\right|_{n \text{ fixed}} \;=\; 1, \tag{C.1}$$

at fixed n, since a constant-elasticity markup is multiplicative and passes both a proportional cost change and a proportional tax through one for one. The amplification in Proposition 6 comes *entirely* from the entry margin, that is from $\mathrm{d}\ln n < 0$, and not from any difference in short-run pass-through between the two instruments. This is the disciplined version of the claim: at fixed market structure the two taxes pass through identically, so the case against the usage tax rests on entry and on targeting, never on an assumed pass-through asymmetry. We prove (C.1) in Appendix A and use it in both propositions.

[C.1] The timing matters for a different reason. Once investment is sunk, taxing its quasi-rents distorts nothing static, and the forward-looking cost is the marginal model or entrant that is then not built. Appendix C.4 sets out both horns of the resulting time-consistency problem (Kydland and Prescott, 1977).

The amplification is large. At $\mu = 1.5$ and $\sigma_q = 1.965$, $\Phi = 1.65$: a usage tax of one percent raises the buyer price of AI by about 1.65 percent, because it also drives out entrants and lets the survivors raise their margins. Under free entry a usage tax on a concentrated sector with heavy fixed costs is therefore not a tax on rent: there is no rent, and its burden on buyers exceeds the revenue it raises.

What survives the obvious caveats is the direction.[C.2] Any model in which fixed costs are large and entry responds to market size delivers $\Phi > 1$, and the static formula therefore *understates* the distortion of a usage tax.

The Union of Concerned Scientists identified 130 transmission projects across seven states built only to connect data centers, and found that more than 95 percent of the connection costs, some $4.3 billion in 2024 alone, were spread across all ratepayers rather than charged to the facilities that caused them (Jacobs, 2025). In the capacity market of the largest U.S. grid operator, the independent market monitor attributed 63 percent of a record price increase to data center load, translating to roughly $9.3 billion in a single delivery year that will be recovered from 65 million customers (Monitoring Analytics, 2025). The pass-through is visible on household bills: residential customers in parts of the mid-Atlantic have seen monthly increases of $15 to $25 attributed to capacity prices, and consumer advocates project cumulative increases of over $100 billion across the region by the early 2030s (Natural Resources Defense Council, 2025).

C.4 Modelling choices and caveats

This appendix collects three qualifications to the model of Section 3 that would otherwise sit in long footnotes.

The grid-cost wedge and the pricing benchmark. The wedge is a statement about marginal cost causation, not about the level at which the utility prices. Regulated utilities recover fixed and capacity costs through average-cost (break-even) tariffs rather than at marginal cost, so the relevant benchmark here is not whether price sits above or below marginal cost, which under a break-even constraint it must on average. What γ isolates is that the *incremental* capacity cost caused by the marginal large load is allocated to other customer classes rather than to the load that causes it; that misallocation survives whether one benchmarks against marginal or average cost. The general tension between

[C.2] Two are worth recording. Free entry can be socially excessive in the sense of Mankiw and Whinston (1986), so a reduction in n is not automatically a welfare loss, and the business-stealing effect works against the amplification result. And the Cournot structure is doing work: with contestable markets or a competitive fringe, $\Phi \to 1$.

recovering fixed network costs and pricing at social marginal cost is the one Borenstein and Bushnell (2022) analyse for residential tariffs; the large-load allocation problem is its counterpart on the industrial side. In a frictionless long-run competitive equilibrium marginal and average cost coincide and the question would not arise, but cost-of-service ratemaking with large point loads is exactly the case in which they do not. Separately, a large new load can also *lower* average rates when it improves utilization of fixed assets that are already in the rate base, since the fixed costs are then divided among more kilowatt-hours (Borenstein, 2025); a detailed study of the largest data center market found no evidence of a historical cross-subsidy and attributed only about half of a recent capacity price spike to load growth (Riu et al., 2025). Both can be true at once.

Fixed-cost recovery, ex ante and ex post. The argument is ex ante, not ex post. Once the fixed investment is sunk, a tax that drives accounting profit negative is indeed non-distortionary on any static margin, and a planner facing an already-built model could tax its rents, and even its fixed-cost recovery, without discouraging an investment that has already happened. Our claim is the forward-looking one: if providers anticipate that fixed-cost-recovery margins will be taxed away, the marginal frontier model, chip cluster, or entrant is not built, which is the margin Section 5.5 makes endogenous. The two readings are the two horns of the standard time-consistency problem (Kydland and Prescott, 1977): taxing sunk capital is tempting precisely because it looks free, and looks free precisely because the commitment not to do so is what elicited the investment. We take the side of the problem in which the government cannot separately promise not to expropriate fixed-cost recovery, so the distinction between rent and fixed-cost recovery has bite; a reader who assumes full commitment and a one-shot, post-investment game will read the fixed-cost-recovery portion as just another taxable rent, and for that reader the market-power term is correspondingly larger.

Scope of the externality. The externality attaches to the electricity drawn by the AI sector, not to final-good production, and not to labor. We are not assuming AI is the only emitting activity in the economy; we are isolating the emissions associated with compute, which is the object of the policy debate this paper addresses. Economy-wide emissions from other sectors are outside the model and are, by construction, the province of the general carbon price τ_c, whose incompleteness ($\delta > \tau_c$) is what leaves a residual wedge on AI's electricity in the first place. This matters for the instrument comparison below: because the wedge lives on the kilowatt-hour delivered to the data center, the

corrective charge t_E of Proposition 3 falls on that node alone and does not touch routine or non-routine labor, whereas a broad carbon price would of course fall on all fuel use in the economy. For the grid-cost component the two are complements. For the carbon component they are substitutes: a higher economy-wide carbon price reduces the top-up the data-center charge must supply.

D Additional calibration detail

This appendix reports the calibration material summarized in Section 6.

D.1 How much rides on σ?

Every magnitude above depends on the elasticity of substitution between AI and routine labor, and σ is the parameter we know least well, so we report how much of the argument survives across its plausible range rather than resting on a single value. The early evidence bounds the range loosely. Controlled deployments find large productivity gains from generative AI concentrated among lower-skilled workers in writing and support tasks (Noy and Zhang, 2023; Brynjolfsson et al., 2025), which is consistent with substitution, though productivity gains alone cannot identify σ: they also arise under complementarity. Our restriction $\sigma > 1$ is therefore a maintained assumption rather than an estimate, and the tasks AI reaches are still expanding. We therefore report every result across $\sigma \in [1.2, 3]$ and let the reader locate their own prior within it.

Figure D.1 carries three messages. First, the predistribution term rises with σ and is not negligible at high elasticities: it is 1.1 percent at $\sigma = 1.2$ and 4.6 percent at $\sigma = 3$, against 9.8 percent at the fixed-assignment corner. Second, the sign of the optimal usage tax is set by ownership and market structure rather than by σ: for an importer with entry blocked it runs from about 4 to 8 percent, and for an economy that owns its providers it is negative throughout, from about -18 to -14 percent. No value of σ reconciles the two. Third, and this is the uncomfortable one, the distributional *stakes* are far more sensitive to σ than the optimal *tax* is. The compression from cheap AI runs from under one log point at $\sigma = 1.2$ to about 42 log points at $\sigma = 3$ measured on group welfare, and two and a half times that on the relative task wage. If σ is high, the appropriate response is a large adjustment to the income tax and transfer system, not a marginal usage tax that works through the same channel as the price movement it is trying to offset.

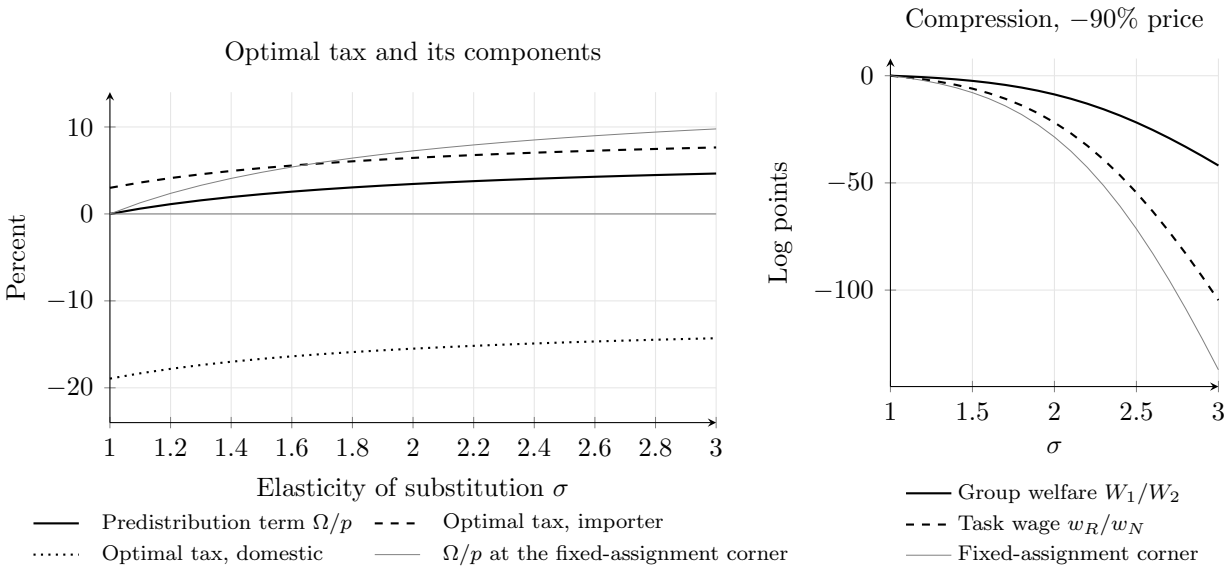


Figure D.1: Sensitivity to the substitution elasticity at the central calibration: $\Delta_\ell = 0.4$, $\kappa = 1.5$, reallocation in efficiency units ($\mathcal{E} = 0.40$), the calibrated earnings base, and $\mu = 1.5$ with entry blocked. The thin grey lines reproduce the fixed-assignment corner in which the robot-tax literature works. Left: the predistribution term and the optimal usage tax for an importing economy and for one that owns its providers. Right: the compression following a ninety percent fall in the price of AI, for the relative task wage and for the relative ex ante group welfare index, which differ by the factor Δ_ℓ. Other parameters as in Table 4.

D.2 Calculation 3: cheap AI and the wage distribution

What cheap AI does to the wage distribution. Now run the model in the other direction, which is where the anxiety actually comes from. Suppose the price of AI falls by ninety percent at constant capability, roughly what has happened over the past two years. With CES and $\sigma = 2$, the odds ratio of the cost shares satisfies $s/(1-s) \propto q^{-1}$, so a tenfold fall in q multiplies the odds by ten. Starting from $s = 0.05$, the AI share within automatable tasks rises to

$$s' \;=\; \frac{10 \times 0.0526}{1 + 10 \times 0.0526} \;\approx\; 0.35.$$

Exposure rises from 1.5 to 10.5 percent of GDP, and by Corollary 1 the labor share falls from 98.5 to 89.5 percent. The routine wage falls relative to the non-routine wage by $\ln\big[(1-s')/(1-s)\big]/(1+\varepsilon_L+\mathcal{E})$, which is -29 log points at $\sigma = 2$ at the fixed-assignment corner but about -22 at $\mathcal{E} = 0.40$, and the gap in ex ante group welfare is Δ_ℓ times that, about -9 log points. As Figure D.1 shows, this is highly convex in σ, running from under one log point at $\sigma = 1.2$ to about -42 at $\sigma = 3$.

This number should be read as an illustrative upper bound, not a forecast, and the reason is a limitation of the two-type structure. Thümmel (2023) works with a routine/non-

routine split, as we do; our departure is to let workers sort across tasks, so that the group a planner weights and the task the technology reaches no longer coincide. Our two-type model collapses that continuum and so delivers the compression as a single relative-wage number, which overstates how cleanly the burden falls on an identifiable group. Lemma 3 makes the same point from the tax side: a shock this large reshapes the earnings distribution in ways a linear income tax cannot neutralize by a proportional adjustment, so the clean separation of common and relative components on which the predistribution term rests is itself strained at these magnitudes. Both caveats point the same way. The calculation establishes that the distributional stakes *can* be large, not that they fall as a tidy transfer between two blocs, and certainly not that a usage tax is the instrument to address them. If a shock of this size arrives, the instrument that meets it is a genuinely nonlinear income tax and the transfer system, not the price of a token, which works through the same weak channel as the price movement it is trying to resist and is an order of magnitude too small to offset it.

D.3 Calculation 5: the welfare cost of the wrong instrument

The welfare cost of choosing the wrong instrument. It is worth putting one number on the central recommendation, since the paper otherwise trades in ratios. Suppose a government wishes to raise revenue equal to 0.05% of GDP from the AI sector, a modest target. Levied as a usage tax, the required rate is about 3.3% on a base of 1.5% of GDP, and by Proposition 6 the quantity distortion is amplified by $\Phi = 2.5$ through the entry margin. The resulting deadweight loss, of Harberger form $\frac{1}{2}\,\Phi\,\sigma_q\,pa\,t_U^2$, is about 0.004% of GDP, or roughly eight cents of deadweight per dollar raised. Levied instead as a cash flow tax on rents, Proposition 4 makes the deadweight loss zero to first order: the same revenue is raised with essentially no distortion, or on the pessimistic residual wedge of §7.2 about two cents on the dollar rather than eight. And levied as an emissions charge, the "cost" is negative, because the instrument corrects a pre-existing externality worth far more per unit than the revenue it raises; the first \$0.05 per kWh of the charge buys a social gain, not a loss. The ranking is therefore not close. Against the rent tax the usage tax loses eight cents on the dollar; against the emissions charge it forgoes a first-order welfare gain to incur a deadweight loss. These are small numbers in levels only because the AI sector is still small; they scale with exposure θ, which every forecast has rising several-fold within the decade, so the cost of entrenching the wrong instrument now grows with the sector.

Table D.1: Probabilistic calibration: the distribution of $t_U^\star$

	Independent priors			Correlated priors		
	Median	5th	95th	Median	5th	95th
Predistribution term	+2.2%	+0.6%	+6.0%	+2.1%	+0.5%	+6.5%
Net externality term	+2.7%	+1.0%	+5.9%	+2.8%	+0.9%	+6.2%
Market power, entry blocked	−20.3%	−42.4%	−9.0%	−20.3%	−44.4%	−8.6%
$t_U^\star$, entry blocked, domestic	−14.8%	−38.3%	−1.4%	−14.7%	−40.5%	−0.6%
$t_U^\star$, entry blocked, importer	+5.4%	+2.1%	+10.0%	+5.5%	+1.9%	+10.6%
$t_U^\star$, free entry	−9.7%	−24.6%	−0.9%	−9.6%	−24.7%	−0.5%
$\Pr(t_U^\star < 0)$, entry blocked, domestic		0.970			0.959	
$\Pr(t_U^\star < 0)$, free entry		0.969			0.960	
$\Pr(t_U^\star > 10\%)$, entry blocked, importer		0.050			0.069	
$\Pr(\Omega/p < 5\%)$		0.901			0.882	

Notes. 200,000 draws from the priors of Section 6.4. The net externality term is D/p before spillovers; the optimal taxes net ψ. The correlated columns impose a Gaussian copula with $\rho(\sigma, \hat{g}_1) = 0.3$, $\rho(d, \text{energy share}) = 0.3$ and $\rho(\kappa, \Delta_\ell) = -0.4$. The free-entry optimum is interpolated from exact equilibrium solutions.

Sensitivity to σ. The substitution elasticity is the parameter we know least well, and Appendix D.1 reports every result across $\sigma \in [1.2, 3]$. Three findings there are worth carrying forward. The predistribution term is bounded: it rises with σ but never leaves single digits anywhere in the range. The sign of the optimal usage tax is set by ownership rather than by technology, at every point in the range. And the distributional stakes are far more sensitive to σ than the optimal tax is, at the calibrated sorting parameters the relative group wage index moves by under half a log point at $\sigma = 1.2$ and by about 28 log points, a fall of roughly a quarter, at $\sigma = 3$, while the tax the model recommends barely moves. That last point is the uncomfortable one: the usage tax is close to irrelevant to the distributional question at exactly the elasticities where that question becomes urgent, which is the model telling us that a large σ calls for the income tax and transfer system rather than a marginal wedge on compute.

D.4 The probabilistic calibration in full

Independence across parameters is itself an assumption. The right panel of Table D.1 reimposes the same marginals through a Gaussian copula with $\rho(\sigma, \hat{g}_1) = 0.3$, because a planner who believes AI substitutes strongly for routine labor is plausibly more concerned about the exposed group; $\rho(d, \text{energy share}) = 0.3$, because carbon-intensive grids tend to serve energy-intensive workloads; and $\rho(\kappa, \Delta_\ell) = -0.4$, because sharp sorting and fast reallocation are substitutes even beyond the deterministic frontier. Correlation fattens the lower tail but leaves the medians close to unchanged, and the probability of a negative

domestic optimum moves only from 0.97 to 0.96 with entry blocked.

Three restrictions do most of the work in locating the distribution: the Lerner bound, which keeps every markup feasible; the shareholder weight derived from $\hat{g}_1$; and the reallocation elasticity computed from the exposure gap. The market-power term exceeds the sum of the other two in 0.97 of draws with entry blocked. The electricity wedge d/p_e exceeds fifty percent in three quarters of draws, which is why the claim that the wedge on the kilowatt-hour is large survives the contested range of the grid-cost parameter. All of these figures assume constant-elasticity demand; Section 6.5 reports what changes when providers absorb part of a tax.

D.5 Pass-through and the optimal tax

Lemma D.1 (Tax–cost equivalence)**.** *With the number of providers fixed and any symmetric conduct, an ad valorem usage tax t_U levied on the producer price affects equilibrium prices and quantities exactly as a proportional increase in every provider's marginal cost from c to $c(1+t_U)$. Hence $\rho \equiv \mathrm{d}\ln q/\mathrm{d}\ln(1+t_U) = \mathrm{d}\ln q/\mathrm{d}\ln c$, the elasticity of the buyer price with respect to an industry-wide proportional shock to marginal cost, and in levels $\rho = \varphi/\mu$, where $\varphi \equiv dq/dc$ is cost pass-through.*

Proof. Provider i receives $q/(1+t_U)$ per unit, so its profit is $[q/(1+t_U)-c]\,Q_i = [q-c(1+t_U)]\,Q_i/(1+t_U)$. Since $1/(1+t_U)$ is a positive constant, every first-order condition, under quantity or price competition, is that of a provider with marginal cost $c(1+t_U)$ and no tax. The buyer price therefore depends on c and t_U only through $c(1+t_U)$, which gives $\mathrm{d}\ln q/\mathrm{d}\ln(1+t_U) = \mathrm{d}\ln q/\mathrm{d}\ln c$. Writing $\mathrm{d}\ln q/\mathrm{d}\ln c = (dq/dc)(c/q) = \varphi/\mu$ gives the levels form. Under constant-elasticity demand the markup is constant and $\rho = 1$; under linear demand with n Cournot providers, $\varphi = n/(n+1)$ and $\rho < 1$. □

Derivation of (10). Hold the number of providers fixed, let the producer price p respond to the tax, and weight buyer surplus and public funds at one and provider profit at $(1-f)g_S$. With $q = (1+t_U)p$,

$$dW = -Q\,dq + d(t_U pQ) + (1-f)g_S\,d\big[(p-m)Q\big] - (\Omega + D - \psi)\,dQ.$$

Using $dq = p\,dt_U + (1+t_U)\,dp$ and $d(t_U pQ) = pQ\,dt_U + t_U(p\,dQ + Q\,dp)$,

$$dW \;=\; -\big[1-(1-f)g_S\big]\,Q\,dp \;+\; \big[t_U p + (1-f)g_S(p-m) - (\Omega + D - \psi)\big]\,dQ.$$

The first term is new: when the producer price falls, part of the tax is borne by providers, and that transfer is worth the gap between public funds and the weight on profit. At $t_U = 0$, write $u = \ln(1 + t_U)$, so that $\mathrm{d}\ln p/du = \rho - 1$ and $\mathrm{d}\ln Q/du = -\sigma_q\rho$. Then

$$\frac{dW}{du} \;=\; pQ\Big\{\big[1-(1-f)g_S\big](1-\rho) + \sigma_q\rho\Big[\frac{\Omega + D - \psi}{p} - (1-f)g_S\Big(1-\frac{1}{\mu}\Big)\Big] - \sigma_q\rho\, t_U\Big\},$$

and setting the bracket to zero gives (10). Under constant-elasticity demand $\rho = 1$, the first term vanishes, and the expression is exact at any rate (Proposition 7); otherwise ρ and σ_q vary with the tax and the expression is first order. Under free entry, profit is identically zero, so the profit terms drop, and Proposition 6 gives $\rho = \Phi$; the expression becomes $(\Omega+D-\psi)/p+(1-\Phi)/(\sigma_q\Phi)$, which is Proposition 8. We checked the expression for dW/du against numerical derivatives of welfare under linear and constant-elasticity demand; it matches to five decimal places.

D.6 Evidence on pass-through

This appendix reports what public data can and cannot say about ρ. By Lemma D.1, ρ equals the elasticity of the buyer price with respect to an industry-wide shock to marginal cost, so the question is whether AI prices move proportionally with cost.

Why closed models cannot be used. Marginal cost per token is the accelerator rental rate divided by throughput, tokens per accelerator-hour. Rental rates are observed: Silicon Data's H100 index records a hyperscaler median of $7.76 per GPU-hour in late 2023, $7.92 in the first half of 2024, $9.34 in the second half and $8.96 in early 2025 (**?**). Throughput is not observed for closed models, whose sizes and serving methods are not disclosed, and it dominates. Using the Epoch Capabilities Index as a measure of capability (**?**), Claude 3 Opus (February 2024, index 126.9) launched at $75 per million output tokens and GPT-4o mini (July 2024, index 126.6) at $0.60 (**?**): a fall of more than ninety-nine percent for the same measured capability, over a period in which the rental rate rose by about eighteen percent. Rental rates explain essentially none of the variation in prices, and a regression of prices on them would have an omitted variable, throughput, many times larger than the regressor. List prices also rarely change within a model version, so even the variation that exists arrives through new models rather than repricing.

Data. Open-weight models solve the cost problem. Their developers sell them through their own APIs, and the same weights are served by independent hosts competing on price, whose prices are a market-based proxy for marginal cost. We use the October 2026 snapshot of a public price tracker that records per-platform prices for open-weight models every six hours, drawing on OpenRouter (**?**). For each model we take the developer's first-party price and the price of every third-party host serving the same weights, one price per host. We exclude hosts that price for latency rather than cost (Cerebras, Groq and SambaNova) and require at least two hosts. The panel has 43 models from nine developers: Alibaba's Qwen (17), Z.ai's GLM (8), MiniMax (4), Xiaomi (4), Moonshot's Kimi (3), DeepSeek (2), Mistral (2), Tencent (2) and Google's Gemma (1). The median model has five hosts, and host prices range from $0.03 to $14.12 per million output tokens.

Specification. We regress the log developer price on the log host price across models,

$$\ln p_m^{\text{dev}} \;=\; \alpha \;+\; \beta \, \ln p_m^{\text{host}} \;+\; \varepsilon_m .$$

Under constant percentage markups, which is the constant-elasticity case, $\beta = 1$; markups that compress as cost rises give $\beta < 1$. Standard errors are clustered by developer.

Results. Table D.2 reports the estimates. Ordinary least squares gives β between 0.64 and 0.81 in every specification and rejects $\beta = 1$. That estimate is biased toward zero, because the host price measures marginal cost with error. We correct for this with a split-sample instrument: for the 30 models with at least four hosts, we split each model's hosts at random into two halves and instrument the median price of one half with the median price of the other, whose measurement errors are independent. Averaging over 2,000 random splits gives $\beta = 0.92$, and a bootstrap over models gives a 95 percent interval of 0.79 to 1.15, with $\beta \geq 1$ in about a fifth of draws. Constant markups cannot be rejected.

What the estimate does and does not show. Three limitations bound the result. First, the market is close to competitive: the developer's price divided by the median host price has a median of 1.00 and an interquartile range of 0.85 to 1.14, so developers have essentially no markup over hosts. Since anyone can serve open weights, that is to be expected, and in a competitive market pass-through near one is expected whatever the curvature of demand. The test therefore says more about the validity of the method than about the closed frontier models, where markups exist and where the market-power term comes from. Second, the variation comes from cost differences across products within a

Table D.2: Elasticity of open-weight developer prices with respect to host prices

Specification	$\hat{\beta}$	s.e.	N
Ordinary least squares			
Output prices, median host	0.756	0.093	43
Output prices, median host, developer fixed effects	0.670	0.121	43
Output prices, cheapest host	0.749	0.088	43
Output prices, cheapest host, developer fixed effects	0.638	0.090	43
Output prices, full-precision hosts only	0.782	0.095	39
Input prices, median host	0.805	0.084	43
Input prices, median host, developer fixed effects	0.807	0.085	43
Correcting for measurement error			
Split-sample IV, output prices (models with ≥ 4 hosts)	0.924	[0.785, 1.150]	30

Notes. Ordinary least squares standard errors are clustered by developer. The split-sample IV estimate averages over 2,000 random splits of each model's hosts; the bracket is a 95 percent bootstrap interval over models. Full-precision hosts serve the model at fp8, bf16 or fp16. Source: tokencanopy price tracker, October 2026 snapshot, OpenRouter data.

single snapshot, not from a common shock over time; Lemma D.1 concerns the latter, so the estimate is evidence on whether percentage markups are constant rather than a direct measure of ρ. Third, the samples are small. The tracker has logged every price change since September 2026, so a panel of within-model price changes will soon allow the second limitation to be addressed with the same data.

Implications. Table 5 evaluates (10) at the values in Table D.2. The domestic optimum stays negative for any ρ above 0.53, which includes every estimate here. The importer's optimum rises from 6 percent at $\rho = 1$ to about 20 percent at the lower bound of the corrected interval and 35 percent at the lowest uncorrected estimate. On present evidence, then, the sign of the domestic wedge is robust to pass-through and the size of the importer's wedge is not.

D.7 Scoring under linear demand

Table 6 scores the instruments under the constant-elasticity demand the model generates. Table D.3 repeats the exercise with entry blocked under linear demand calibrated to the same markup, $\mu = 1.5$, and the same demand elasticity at the initial equilibrium, so that providers absorb about forty percent of a cost increase. Every sign that depends on pass-through reverses. The usage tax and the DST raise welfare for a domestically owned sector, slightly, and for an importer substantially, because a tax that providers partly absorb extracts rent that the planner values below public funds or not at all. The emissions charge raises an importer's welfare for the same reason. What survives the

Table D.3: Welfare change from each instrument, constant-elasticity and linear demand (% GDP)

		Domestic, entry blocked		Importer, entry blocked	
Instrument	Revenue (linear)	CES	Linear	CES	Linear
Token / usage tax, 10%	0.131	−0.052	+0.005	+0.004	+0.080
DST-style levy, 3%	0.043	−0.014	+0.003	+0.004	+0.028
Emissions charge, d per kWh	0.043	−0.029	−0.003	−0.015	+0.020

Notes. Central calibration, $\mu = 1.5$, $(\Omega + D)/p = 6.44\%$, $g_S = 0.66$. Linear inverse demand with Cournot conduct, calibrated to the same markup and to an elasticity of 1.965 at the initial equilibrium. The emissions charge is scored on the scale margin only.

change in demand curvature is the ordering by ownership.

D.8 Parameter provenance and the electricity wedge

This appendix collects the provenance table referenced in Section D.8 and the robustness of the electricity wedge to the contested grid-shifting parameter γ.

Table D.4: Parameter provenance

Par.	Value / range	Evidence	Design	External-validity caveat
σ	2; $[1.2, 3]$	Gains concentrated among lower-skilled workers in writing, support (Noy and Zhang, 2023; Brynjolfsson et al., 2025)	RCT; firm rollout	Task gains are not a production elasticity; rules out $\sigma \leq 1$ but pins no point estimate.
$\hat{g}_1$	0.3; $[0.1, 0.5]$	Normative	—	A social preference, not an estimate; reported across its range.
ε_L	0.3; $[0.1, 0.5]$	Standard range in public finance	Quasi-experimental	Intensive-margin estimates; the margin here may be occupational switching.
κ	1.5; $(1, 3]$	Roy–Fréchet dispersion of comparative advantage (Hsieh et al., 2019; Galle et al., 2023)	Structural Roy model	Estimated for occupational and sectoral sorting, not for AI exposure. Enters reallocation as $\kappa - 1$ in efficiency units; $\kappa > 1$ for a finite expected wage.

Table D.4 (continued)

Par.	Value / range	Evidence	Design	External-validity caveat
Δ_ℓ	0.4; measured $[-0.50, +0.45]$	*Computed here* (Table 2, Appendix D.9) from O*NET 30.3 education distributions (National Center for O*NET Development, 2026) and seven public exposure measures (Webb, 2020; Felten et al., 2021; Eloundou et al., 2024; Eisfeldt et al., 2023), employment-weighted with Dorn (2009) labor-supply weights	Descriptive; occupation-level	Set to one by assumption throughout the robot-tax literature; that corner is empirically false and quintuples the term. The carried 0.4 is the value on the single most favourable measure; all four direct AI measures give a *negative* gap (Remark 2). Occupation-level education distributions substitute for ACS microdata, so the within-occupation grouping is imputed.
χ	≈ 0.5	*Derived*, not calibrated: share of the textbook correction absorbed by free entry for the ad valorem tax (Table B.1)	Theory	Demand-system dependent. Not a free parameter, and not used with entry blocked.
g_S	0.66; $[0.45, 0.85]$	Implied by $\hat{g}_1 = 0.3$ under isoelastic weights, given the earnings position of routine occupations and the concentration of equity ownership (Lemma 4)	Calibration	Depends on the isoelastic form and on log-normal earnings; frontier-AI equity is likely more concentrated, lowering g_S.
ϕ	$[0.2, 0.4]$	Recurring training and capacity investment as a share of revenue	Accounting	Not separately observed; shrinks the cash-flow base and indicates which regime applies (§B.2). Does not scale the market-power term.
μ	1.5; $[1.2, 2.0]$	Within the Lerner bound $1 - 1/\mu \leq 1/\sigma_q$; reported gross inference margins of 50–60 percent	Company disclosure; theory	Margins at the bound imply near-monopoly conduct; parametrized by conduct $h = \sigma_q(1 - 1/\mu)$.
Group-1 earnings share	0.45; $[0.40, 0.50]$	Share of labor income earned by the group the planner weights	Assumption	Sets the earnings base β_1 of the predistribution term; equals the routine share only under fixed assignment.
Pass-through	1 at fixed n	Implied by constant-elasticity demand	Theory	**Not measured for AI.** Below one, the sign of the domestic optimum reverses (§6.5).
ζ	0.4 kg/kWh	Grid intensity serving U.S. data centers (International Energy Agency, 2025)	Measurement	Varies by an order of magnitude across sites; a fleet average hides the siting margin.
δ	\$190/t	Social cost of carbon (U.S. Environmental Protection Agency, 2023)	Integrated assessment	Generous: above most market prices and above Nordhaus (2017).

Table D.4 (continued)

Par.	Value / range	Evidence	Design	External-validity caveat
γ	$[0, 0.03]$ /kWh	Socialized capacity and transmission cost (Jacobs, 2025; Monitoring Analytics, 2025) vs. (Energy and Environmental Economics (E3), 2026; Riu et al., 2025)	Cost allocation	*Genuinely disputed.* Figure D.2: the level conclusion survives the range; §6.6 is where it bites.
e	0.3–150 Wh	Energy per query by workload (International Energy Agency, 2026; Luccioni et al., 2024; de Vries, 2023)	Benchmark	Workload mix is not publicly observed; see notes to Table 1.
θ	1.5%	AI spending as a share of output	National accounts	Generous; rising fast, so a statutory rate dates quickly.
η_E	0.1 SR; ≈ 1 LR	Price elasticity of U.S. electricity demand (Burke and Abayasekara, 2018); meta-analysis (Labandeira et al., 2017)	IV panel, state $\times$ sector $\times$ year	Not estimated for data centers. Much of the large long-run industrial elasticity reflects clustering in low-price states, leakage ℓ here, not abatement.
ℓ	**unmeasured**	—	—	**No credible estimate.** Proposition B.1, the one surviving case for a usage tax, rests on it.
ψ	**unsigned**	—	—	**Not signed by existing work.** Spillovers push up, business stealing down; any $\psi > 0$ lowers $t^\star_U$.

D.9 The exposure gap: estimand and construction

Because Δ_ℓ is new to this literature and is the input on which the predistribution magnitude chiefly depends, we state it precisely enough to be computed and disputed.

Estimand. Fix a partition of the employed workforce into two ex ante groups and a binary classification of occupations into exposed and unexposed. Then

$$\Delta_\ell \;=\; \ell_{1R} - \ell_{2R}, \qquad \ell_{iR} \;=\; \frac{\sum_{n\in i} \omega_n \, \mathbf{1}\{o(n) \in R\}}{\sum_{n\in i} \omega_n},$$

with ω_n the survey weight and $o(n)$ the occupation of worker n. Three features are worth stating because each is a place the literature's convention hides an assumption. First, $\Delta_\ell \in [-1, 1]$: it is signed, and Remark 2 explains why the sign is not guaranteed. Second, it is a difference of shares rather than a regression coefficient, so it has no residual variance

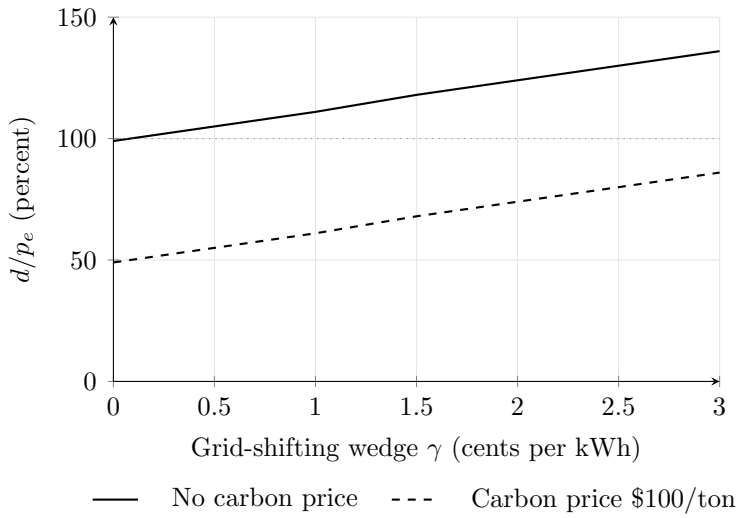


Figure D.2: The electricity wedge is large for any γ. The residual external cost per kilowatt-hour, as a share of the price a data center pays, plotted against the socialized grid-cost wedge γ, with and without a pre-existing carbon price. Even at $\gamma = 0$ the wedge is roughly half to all of the electricity price, because unpriced carbon dominates. Across its contested range γ moves the wedge by up to about 37 percentage points, less than carbon does; its larger role is distributional. The carbon-price line uses $100 per ton, the offset implied by the central d. Other parameters as in Table 4.

and needs no identifying assumption beyond the classification itself; the only sampling uncertainty is that of two proportions. Third, it is defined on *ex ante* groups, which by Assumption 1 are unobservable, so any implementation replaces them with an observable proxy and Δ_ℓ is therefore proxy-dependent. That is a feature: the planner's instrument set is also proxy-dependent, and the relevant Δ_ℓ is the one attached to whatever grouping the planner's welfare weights actually track.

Construction. One year of American Community Survey microdata, restricted to employed workers aged 25 to 64, merged by SOC code to an occupation-level exposure score, with "exposed" defined as a score above an employment-weighted quantile. Three choices have to be reported rather than fixed:

- *Exposure measure.* At minimum the routine-task intensity of Autor and Dorn (2013), with the conventional cut at the 66th percentile, and the LLM exposure score of Eloundou et al. (2024), cut at the median. Table 2 uses the robot and software measures of Webb (2020) on the routine side in place of routine-task intensity, which remains to be computed. The first matches the object σ describes; the second matches what the technology currently does. They disagree in sign, which is the substantive finding rather than a nuisance.
- *Ex ante grouping.* Less than a bachelor's degree versus the rest; high school or

less versus the rest; below versus above median earnings; and a young non-college interaction. The paper's argument requires only that the maximum of $|\Delta_\ell|$ across these be well below one.

- *Threshold.* The share of employment classified as exposed. Reporting Δ_ℓ across a range of thresholds is cheap and is the natural robustness check, provided one remembers that gaps fall mechanically as the cut tightens, so comparisons across thresholds should use the gap relative to its attainable maximum.

Implementation. Table 2 implements the occupation-level counterpart of the estimand above. Absent worker microdata, group membership enters as a probability rather than an indicator: writing E_o for employment in occupation o and $p_o = \Pr(\text{group } 1 \mid o)$,

$$\ell_{1R} = \frac{\sum_o E_o\, p_o\, \mathbf{1}\{o \in R\}}{\sum_o E_o\, p_o}, \qquad \ell_{2R} = \frac{\sum_o E_o\, (1 - p_o)\, \mathbf{1}\{o \in R\}}{\sum_o E_o\, (1 - p_o)}.$$

This coincides with the microdata estimand whenever exposure is assigned at the occupation level, which it is for every published measure, and it differs from it only through the accuracy of p_o. We take p_o from the incumbent-reported Required Level of Education distribution in O*NET 30.3 (National Center for O*NET Development, 2026), averaged from eight-digit to six-digit SOC and renormalized to sum to one across the twelve education categories; "less than BA" is categories one to five, "HS or less" categories one and two, "less than associate" categories one to four. The fourth grouping uses the O*NET job zone, a preparation rather than an education measure, and is included because it is the closest available substitute for the earnings grouping that microdata would permit. Employment weights are the Dorn (2009) `occ1990dd` labor-supply weights from the replication file of Webb (2020), allocated to detailed SOC occupations: exactly at the level of the SOC major group, and within a major group in proportion to a title match between the harmonized 1990 census categories and SOC titles. The resulting panel covers 765 detailed occupations and 152 million workers, with a major-group employment distribution close to the published national one. The Felten–Raj–Seamans and Eisfeldt et al. scores are carried from SOC 2010 to SOC 2018 with the BLS crosswalk (U.S. Bureau of Labor Statistics, 2018); the Eloundou et al. scores are averaged from O*NET-SOC to six-digit SOC.

Robustness. Three checks. Dropping the employment weights entirely and treating each detailed occupation equally leaves the robot gap at +0.44 against +0.41 and moves

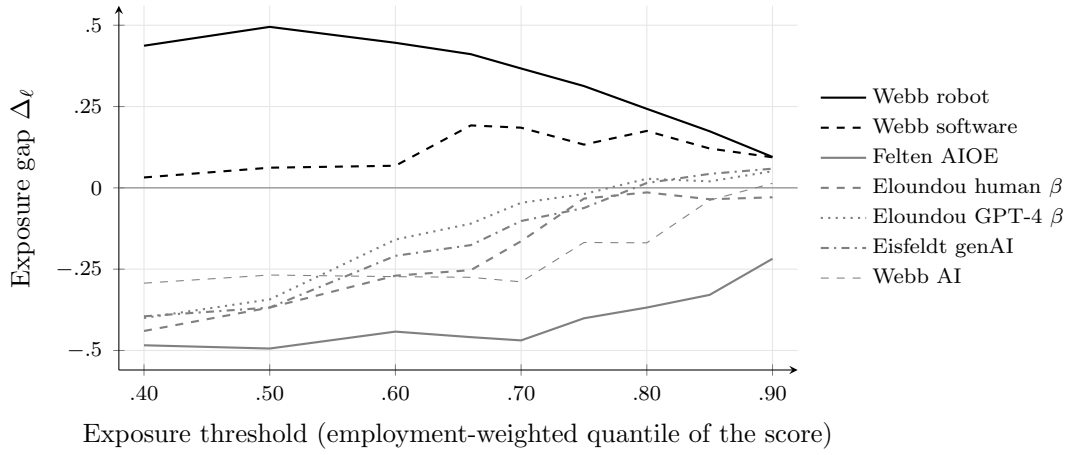


Figure D.3: The exposure gap as the definition of "exposed" is tightened, at the less-than-BA grouping. Black lines are the two routine-displacement measures, grey the four direct AI measures. Six of the seven fall toward zero, and at the ninetieth percentile every gap lies between −0.22 and +0.10. The fall is partly mechanical, since the gap is bounded by $\min\{E/\pi, (1-E)/(1-\pi)\}$ for exposed share E and group share π, so it does not by itself measure segregation. The seventh, Webb software, never leaves a narrow band around +0.1 at any threshold. The vertical scale is the same as the horizontal scale of the left panel of Figure 1, and both stop well short of the value the literature assumes.

every AI measure further negative, to between −0.49 and −0.61. The weighting is not what produces the sign pattern. We weight because detailed occupations vary enormously in size and the boundaries between them are partly conventional, so an unweighted average answers a question about the classification and not about the workforce. Tomlinson et al. (2025) find the same choice mattering in the same direction: their wage gradient is weaker once occupations are weighted by employment, which they attribute in part to earlier work not weighting. Their measure of what AI is actually used for correlates with occupational wages at +0.13 weighted and +0.17 to +0.32 unweighted, and they attribute part of the strong positive wage gradient in earlier work to the absence of weighting. Our sign pattern survives either way, which is why we report both. Varying the exposure threshold compresses every gap toward zero except the already-small Webb software gap, which is flat; Figure D.3 plots the full sweep. And the four groupings never disagree in sign on any measure, with a maximum spread of 0.09 within a row, so the result is not an artifact of where the education line is drawn.

Limitations. We see three, and together they are a reason to read the numbers as descriptive evidence about the chosen proxies rather than as point estimates; none of them biases the gap in a known direction. First, p_o is the education distribution that O*NET incumbents report as *required*, which is not the distribution actually held. It is probably more concentrated than the truth, and concentration makes groups look better sorted across occupations than they are. Second, the allocation of employment within

a SOC major group rests on a title match and not on a validated crosswalk.[D.1] Third, we report no standard errors. The sampling variance of the estimand is that of two proportions in microdata we do not have. The dispersion across the seven measures is an order of magnitude larger than any plausible sampling error, and that is the uncertainty that matters here.

Exposure is not adoption. A fourth issue is conceptual, and it applies to every measure in Table 2 and not to our construction of them. The object σ describes is AI substituting for labor in production. The seven measures describe the scope for it to do so. Bick et al. (2026) compare the two using a nationally representative survey that asks workers which of their own tasks they use generative AI for, and the gap is large. Exposure scores correlate positively with realized adoption, but they account for at most about half of the variation across workers, and occupation-level adoption rates more than double the explanatory power of the leading exposure measures in predicting whether a given worker uses the technology. Much of the residual is individual and not occupational: within narrowly defined tasks adoption is far from universal, and worker fixed effects absorb more variation than task characteristics do.

We expect this to reinforce the pattern in Table 2 and not to overturn it. Their high-adoption occupations are computing, business, finance and management, and their low-adoption occupations are personal service, transport and manual work. This is the gradient that produces our negative gaps, so recomputing Δ_ℓ on adoption should give something at least as negative as the direct AI measures give. We say expect because the occupation-level adoption index is not yet public.[D.2]

Yin et al. (2026) replicate the Eloundou et al. (2024) rating procedure across models and find the resulting exposure scores less stable than their widespread use assumes.[D.3] We therefore read Table 2 as a statement about the spread across measures, not about any single one.

What would change our conclusions. A value of Δ_ℓ above roughly 0.8 on the grouping a planner would plausibly weight would restore the fixed-assignment magnitudes of the predistribution term. Even then, with entry blocked, the domestic wedge stays negative at any markup above about eighteen percent, so the measurement bears more on the size

[D.1] The unweighted check above shows that the sign pattern does not depend on it.

[D.2] It is promised in the paper's replication package, and we would run it against this table when it appears.

[D.3] Two of our seven rows come from that procedure.

of the redistributive case than on the sign of the tax. No measure, grouping or threshold we examined gives a value above 0.51 in absolute value; the largest positive value is 0.50, for robot exposure at the median cut. The falsification is attempted and fails. A value near zero or below removes the predistribution term entirely or reverses it, and that is what the four direct AI measures deliver. The case in the text is the intermediate one, and it is the one favourable to the instrument.

D.10 The CES correction to the exposure cancellation

Lemma D.2 (CES outer nest)**.** *Replace the Cobb–Douglas aggregator in* (1) *with a constant-elasticity-of-substitution aggregator,* $Y = \Lambda\big[\lambda_0 X_A^{(\eta-1)/\eta} + (1-\lambda_0) X_N^{(\eta-1)/\eta}\big]^{\eta/(\eta-1)}$*, in which* η *is the elasticity of substitution between the automatable and non-automatable bundles and* $\eta = 1$ *recovers the case of the text. Then, to first order in* s*,*

$$\frac{\mathrm{d}\ln(w_R/w_N)}{\mathrm{d}\ln q} \;=\; \frac{s(\sigma-\eta)}{\eta+\varepsilon_L+\mathcal{E}},$$

and the predistribution term of Proposition 2 becomes $(1-\tau)\hat{g}_1\Delta_\ell(\sigma-\eta)\beta_1/[(\eta+\varepsilon_L+\mathcal{E})\sigma_q]$*. The exposure* s *still cancels, but the wage channel depends on* σ *relative to* η *and vanishes at* $\sigma = \eta$*.*

Lemma D.2 disarms the obvious objection that the clean cancellation of exposure is an artifact of the unit elasticity. It is not: the cancellation survives to first order for any outer elasticity, and the correction is bounded by exposure, which is small. We use Cobb–Douglas in the body for readability and rely on the lemma for generality.

E Implementation and the policy debate

This appendix collects the implementation detail behind Section 7.

E.1 Administrability of each instrument

Optimal tax formulas assume the base is observable. Each instrument here faces a different administrative constraint, and in one case the constraint is severe enough to qualify a result we have otherwise defended. We follow Keen and Slemrod (2017) in treating administrability as part of instrument choice rather than as an afterthought.

The emissions charge is the easy case. The base is metered by the utility, at the point of delivery, for billing purposes that already exist. Tons of CO_2 follow from kilowatt-hours and a grid intensity factor that system operators already publish. There is no new reporting entity and no transfer-pricing margin, because electricity delivered to a physical site cannot be booked in another jurisdiction. One valuation question does remain: a published grid-average factor need not equal the marginal emissions an additional load causes, so the factor should track the spatial and temporal emissions response, and credit for power-purchase agreements should depend on evidence of additional abatement. This is close to the ideal case for a Pigouvian levy and it is one more reason to prefer the meter.

The rent tax is administratively demanding but familiar. Neutrality requires a cash flow base with immediate expensing and full loss offset. Its advantage is precisely that the authority never has to separate rent from the normal return, since expensing exempts the latter in present value (Auerbach and Devereux, 2018); the difficulties lie instead in implementing full loss offset, allocating shared costs, and taxing nonresident providers. The last is binding for an importing economy: the zero domestic cost of the cash flow tax in Table 6 assumes the jurisdiction can establish taxing nexus over foreign providers and apportion their cross-border revenues and costs, which current rules do not guarantee. The known failure mode is political rather than technical: what gets legislated is usually a levy on gross revenue rather than on cash flow, with the consequences §5.2 describes. The administrative recommendation is therefore specific and checkable. Expensing, loss carry-forward with interest, and a base defined on cash flow rather than turnover.

The usage tax is easy to administer and that is part of the problem. Providers already meter tokens and issue invoices, so the base is available at near-zero cost. Administrative convenience is precisely why the instrument is attractive to legislatures and precisely what the rest of this paper argues against; the point is worth making explicitly, because in a Keen–Slemrod framework a low enforcement cost is a genuine advantage that must be weighed, not dismissed. Our claim is that it does not come close to offsetting a base that is wrong by an order of magnitude on the harm-to-value ratio.

The border adjustment is the hard case, and we do not want to overclaim it. Proposition B.1 is the one surviving argument for a usage tax, and it is also the least administrable thing in the paper. Implementing it requires attributing embodied energy and grid intensity to an imported AI service, which means knowing which data center served a given query and on what grid, information that is not on the invoice, that

providers route dynamically for reasons of latency and capacity, and that no customs authority currently has any means of verifying. Three fallbacks exist, in descending order of fidelity: provider-level attestation of fleet-average energy and carbon intensity, audited as financial statements are; a default benchmark rate applied absent attestation, set conservatively so that attestation is the preferred option; or a rebuttable presumption keyed to the provider's jurisdiction of primary compute. Each is cruder than the emissions charge it substitutes for. This is a genuine limitation of the border instrument, and it is a further argument for the central recommendation: reaching the externality at source, where it is metered, is administratively as well as allocatively superior to reaching it at the border, where it must be inferred.

E.2 The political economy of instrument choice

If the usage tax is so poorly suited to every stated goal, why is it the instrument most often proposed? The model contains the beginning of an answer, and it comes from the entry result rather than from any appeal to confusion among policymakers.

Proposition E.1 (Incumbents prefer the usage tax)**.** *Consider the incumbent providers active before a tax is imposed, and compare a usage tax with a revenue-equivalent cash flow tax on rents. The cash flow tax lowers incumbent profit one for one with the revenue it raises. The usage tax, by Proposition 6, raises the equilibrium markup through the entry channel; incumbents therefore recover a share* $\Phi - 1$ *of the statutory burden through higher margins on retained sales, and for a sufficiently high fixed cost the surviving incumbents can be left* better off *than under no tax at all. Incumbents rank the instruments in the opposite order to the planner.*

The proof is immediate from the pass-through and entry results and is given in Appendix A. The content is a positive prediction, not a normative claim: the instrument that is worst for welfare is the one that the politically most concentrated interest, the incumbent frontier labs, has least reason to oppose and some reason to favour, because it raises rivals' entry costs and lifts margins on retained sales. A cash flow tax falls squarely on those same incumbents with no offset. Standard theories of capture then predict a pattern we have not tested: firmer resistance to profit taxes and competition policy than to usage levies.[E.1]

[E.1] Two further forces push the same way. Revenue salience favours the usage tax, since a fast-growing visible transaction base is easier to legislate and score than economic rent, whose base is contested and whose yield is uncertain. And host districts may resist the emissions charge if it slows investment,

This yields a testable implication that separates the account from mere pessimism. If incumbents prefer the usage tax for the entry-barrier reason, and if firms differ in size and sunk cost, which our symmetric model does not capture, their revealed positions should track those differences: the largest labs least opposed to usage levies and most opposed to rent taxes and interoperability mandates, smaller entrants holding the reverse ranking. That is a pattern one can look for in lobbying disclosures and public comments. The normative implication is not defeatism but sequence: because the worst instrument is the path of least resistance, the reforms most worth defending are the ones the political system will fight hardest.

E.3 Reading the debate

Section 7 scores the four instruments against each other on revenue, distortion and incidence, and we do not repeat that ranking here. Two features of the debate are not captured by any scoring table, and this subsection records them.

Electricity and the grid. This is where the model puts nearly all of the weight, and where the instruments partly exist already. The reforms that matter are not new taxes: they are cost allocation rules that charge large loads for the capacity and transmission they cause, tariffs with meaningful minimum take and exit provisions to protect ratepayers against stranded investment, and a carbon price that applies to data center load like any other. The evidence on whether cost shifting has occurred is genuinely mixed (Energy and Environmental Economics (E3), 2026; Wiser et al., 2026; Joint Legislative Audit and Review Commission, 2024), but the direction of reform does not depend on resolving it. Charging a load for the cost it causes is right whether or not it has so far failed to be charged.

Neutrality. Acemoglu et al. (2020) show that the tax code already tilts the user cost toward capital and away from labor. In the language of (9), that is an existing negative t_U that nobody legislated. The first-best response is to remove it. Where removal is unavailable, an offsetting levy on the same margin reduces the distortion, and the statutory rate that implements a target wedge $t_U^\star$ is then roughly $t_U^\star - t_U^0$, where $t_U^0 < 0$ is the existing implicit subsidy. We do not establish that the capital bias acts on the AI-purchase margin

employment or fiscal revenue there while the climate benefit is diffuse. The direct price burden, by contrast, is spread across AI buyers, and correcting the grid-cost wedge can benefit local ratepayers, so the net local burden is uncertain.

or quantify t_U^0, so its effect on the statutory-rate comparisons is left open.

Remark E.1 (The formula is not stable). The ratio D/p moves over time and in both directions. The price of a token at fixed capability has fallen by roughly an order of magnitude over the past two years, which pushes D/p up. Energy per token has also been falling, which pushes it down. The two trends are measured on different bases, so their net effect on D/p cannot be read off directly. Reasoning and agentic workloads consume far more energy per unit of delivered value, which pushes it up again. A rate written into statute today will be wrong within two years. This is an argument for indexing any levy to a measured physical quantity, kilowatt-hours or tons of CO_2, rather than to the value of a transaction. It is the targeting principle, arriving from a different direction.