Leading Firms and the Future of Work*

Max Brès, Fabrizio Colella, Philipp Kircher and David Koll

November 15, 2026

Most recent version

Please Do Not Cite Without Permission

Abstract

Understanding how occupations will rise or fall with technological and organizational change over the next decade is essential for labor market adjustment. This paper develops an approach to forecasting occupational change by identifying *leading firms* whose workforce composition anticipates the future occupational structure of their industry as other firms catch up over time. We develop a theoretical model with technology diffusion and frontier-laggard dynamics showing that leading firms are early adopters of technology and how this can be exploited for forecasting. We apply its insights by (1) machine-learning current characteristics of firms whose current occupational structure is close to the future occupational structure of its industry, and (2) estimating a forecasting model that combines current and lagged occupational structures of industries and leaders. Application in a hold-out sample a decade later shows substantial improvements in predictions relative to a model without leaders. We outperform official forecasts and predictions based on automation indices. The machine-learned leader forecast for 2023 correlates negatively with several measures of automation–a relationship with reverses for early automation measures when projecting out to 2033, while staying negative for recent AI- and GPT-based measures.

Keywords: Occupational change; leading firms; future of work; labor demand; occupational forecasting; machine learning; technology diffusion; firm heterogeneity

JEL: C53; C55; E24; J21; J23; J62; O33

*Brès: Sveriges Riksbank, max.bres@riksbank.se. Colella: USI Lugano and Rockwool Foundation Berlin, Fabrizio.colella@usi.ch. Kircher: Cornell University and NBER, pk532@cornell.edu. Koll: International Monetary Fund, dkoll@imf.org. We published a pre-analysis plan for this project, which can be accessed at https://osf.io/5gtsw/overview. Nicola Robatto and Leonardo Peñaloza-Pacheco provided outstanding research assistance. Vinzenz Ziesemer made invaluable contributions to an earlier version of this project. The authors are grateful for feedback of seminar participants at Cornell, Louvain-la-Neuve, IMF, Ski & Labor Workshop Laax, AIEL Milan, Stanford SITE, and Arizona State University. The authors gratefully acknowledge funding through the European Research Council Grant No. 818859, and Kircher thanks the Universite Catholique de Louvain where he conducted much of this work. Access to confidential data, on which this work is based, has been made possible within a secure environment offered by Centre d'accès sécurisé aux données (Ref. 10.34724/CASD). **Disclaimer:** The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of Sveriges Riksbank and of the International Monetary Fund, its Executive Board, or its management.

1 Introduction

Waves of technological and organizational change – including skill-biased technological change, computerization and automation, the adoption of industrial robots and artificial intelligence, and the offshoring of production and services - have reshaped the tasks performed within firms and, consequently, the occupations and skills they demand. Forecasting these changes is important for education and retraining: decisions about which skills to acquire or teach must often be made years before the resulting shifts in labor demand become apparent.

The literature has yet to establish a comprehensive method for forecasting such labor market change due to several challenges. One seminal approach – reviewed below – has been to ask experts or analyze patents to determine which tasks and associated occupations could feasibly be automated (e.g. Frey and Osborne, 2017; Nedelkoska and Quintini, 2018; Webb, 2020; Kogan et al., 2021; Martinez and Moen-Vorum, 2024). The challenge is that actual technology adoption does not necessarily track technological feasibility closely, either because the costs of the technology exceed the labor cost savings or because adjustment costs are high (Lindenlaub et al., 2026). Most of these approaches focus more on tasks that can be substituted by technology, and less so on tasks complementary with technology for which demand will rise. A second approach relies on actual technology adoption data, but such data are often hard to obtain. Moreover, defining "adoption" is not straightforward when superficial adoption is cheap but meaningful technological change may require investment into a combination of multiple technological and organizational innovations (Bresnahan et al., 2002; Brynjolfsson et al., 2021). For example, in 2000, more than 80% of large establishments reported basic Internet participation, but fewer than 15% used it for "enhancement," involving changes to internal operations or the introduction of new services (Forman et al., 2003).[1] Superficial adoption is an issue that also arises in more recent technology waves.[2]

This paper proposes a complementary approach by introducing an algorithm to forecast the labor market's occupational structure based on the identification of leading firms. By *leading firms*, we mean firms whose occupational structure is already ahead of the current distribution in their industry. Identifying these firms allows us to extrapolate the aggregate labor market's occupational structure from those that adjust first. We propose a method that identifies leading firms without the need to have data on the exact nature of the underlying technology adoption,

[1] Enhancement in Forman et al. (2003) is identified through a commitment to advanced Internet applications, including Internet-based enterprise resource planning and TCP/IP-based applications in customer service, education, extranets, publications, purchasing, or technical support. The authors use a threshold of at least two such projects and explicitly note that enhancement activities typically also involve changes in organizational procedures.

[2] Bonney et al. (2026) document substantial differences between worker-level AI use and firm-level deployment across business functions, while Bick et al. (2026b) find that generative AI adoption is widespread across occupations but remains relatively shallow at the task level. More generally, reported adoption rates depend strongly on whether surveys measure any AI use or its application in production (Bick et al., 2026a).

show that it improves forecast accuracy considerably in a hold-out sample, and use it to project a decade forward. Going forward, leading firms turn out to employ less labor in occupation with high exposure indexes for AI and large language models, but more in occupations assessed a decade earlier as highly prone to technological automation.

The idea that innovations spread gradually from early adopters to the broader economy has a long intellectual history. Technology diffusion patterns have been documented at least since the influential work of Griliches (1957), Mansfield (1961), and Rogers (1962), and recent evidence confirms that adoption lags remain quantitatively large across technologies and countries (Comin and Hobijn, 2010). These lags imply that at any given time, a subset of firms has already adjusted its production and workforce to technologies that the majority of firms in the same industry have not yet fully adopted. The rise of information technology in the late 1990s illustrates the mechanism: firms such as Cisco and Dell reorganized their workforce around ICT-based supply chain coordination years before their competitors. A similar pattern is emerging with artificial intelligence, where some firms have already automated substantial parts of routine operations, while most have barely begun to expand it beyond individual workers' occasional usage.

The literature on adoption dynamics suggests that there might be firm characteristics that generally define early adopters, across a range of different adoption episodes spanning multiple decades and various technologies. Early adopters tend to be larger, more productive, and pay higher wages (Doms et al., 1997; Koch et al., 2021; Acemoglu et al., 2022; Hampole et al., 2025), and several of these differences are present already *before* adoption (Doms et al., 1997; Koch et al., 2021). This offers the prospect that one can define early adopters at least with some precision by using general characteristics of these firms, rather than needing to specify exactly what combination of technologies constitutes the newest wave of adoption. We innovate on this idea by proposing to machine-learn the characteristics of these leaders: If it is true that these "leaders" foreshadow the occupational structure of their industry, we can use a past episode (in our case the early 2000s) to classify each firm by how close its internal occupational structure is to the future structure of its industry (in the early 2010s). Then we consider a rich set of underlying observables and use a random forest to learn which firm characteristics are associated with being a leader in occupational structure.

When moving to a future period (our hold-out sample in the later 2010s), we classify firms as likely leaders if they have the characteristics of earlier leaders. We also adopt the earlier "catch-up" dynamics of the whole industry to the occupational structure of the leaders to form a forecast of the industry occupational structure (for the 2020s) that we can compare to the actual evolution of the occupational structure. Our illustrative theory outlines both the functional form we use for forecasting as well as the underlying assumptions that make our leader-follower approach feasible: it requires a similar cost-benefit ratio and size ratios of the new technology compared

to previous waves.

This also shows the limitations of our approach and its complementary nature to other approaches: We identify future consequences only of technologies that are already used by at least some firms; it requires some consistency in the cost-benefits of adoption over time; and it requires low enough "noise" in the data so that those firms who foreshadow the future occupational structure are indeed leaders rather than coincidence. If the costs of deep (rather than superficial) technology adoption are partially due to the costs of intangible capital in the form of managerial and organizational changes (see, e.g., Brynjolfsson et al., 2021), it might be more reasonable to presume that the costs of adoption remain somewhat similar over time. And while our approach does not capture future technologies, it has the upshot that it does not have to specify which combination of technologies and managerial/organizational practices currently reflects the frontier, as long as the leading firms have worked it out themselves.

We investigate our approach on matched employer-employee and balance sheet data covering the universe of private sector firms in France from 2002 to 2023. We estimate the model on 2002–2012 and evaluate it out-of-sample on 2013–2023, a period and set of firms not used in estimation. The out-of-sample evaluation follows the protocol of a pre-analysis plan registered before we first accessed any testing data, which we adjusted only to accommodate data quality issues before 2002.[3] We split two digit industries at the median firm size, allowing the identity of industry leaders to differ between small and large firms. In practice, this distinction has little effect on forecasting performance, as shown in the robustness analysis. Our baseline sample includes firms with more than twenty employees at least once over the sample, ensuring that firms are large enough for their occupational structure to plausibly approximate that of the broader industry. We assess the sensitivity of our results to including smaller firms in the robustness analysis. The procedure then proceeds in three steps.

Industry leader characteristics: The random forest identifies the following variables as most influential in characterizing firms in 2002 to 2004 whose occupational structure foreshadows that of their industry a decade later: productivity, industry, employment, sales per worker, average wages, and tangible assets. Interestingly, these characteristics broadly resemble those highlighted in the literature on industry dynamics discussed above, even though we do not train the model directly on measures of technology adoption. We use the resulting model to score firms in both 2004 and 2015 and, within each industry, classify the top 20 percent as leaders, using employment weights. According to our theory the exact size of the leaders should not matter,

[3]The pre-analysis plan can be accessed through this link: https://osf.io/5gtsw/overview. The plan registered the leader identification algorithms, the forecasting equations, and the Theil ratio assessment, with estimation on 1997–2006 and evaluation on 2010–2019 at the three-digit industry level. We subsequently learned of an entire redesign of the DADS processing chain that was introduced between 2001 and 2002, implying that "the data are not directly comparable to previous years" INSEE (2012) INSEE (2013), so we instead split the remaining 2002-2023 period into two equal parts for training and out-of-sample assessment and proceed at 2-digit industry level (initially specified as robustness) to span an industry change in 2008 (see Section 3).

and that is confirmed in robustness checks. Although the random forest allows these characteristics to interact in flexible and potentially complex ways, the firms it identifies as leaders are, on average, larger, more productive, and higher paying than other firms in the same industry. Our forecasting exercise assumes that the characteristics associated with leadership are stable over time. We assess this assumption retrospectively by comparing the leaders selected using the 2004 model with the firms that would have been identified as leaders in 2014 had the subsequent evolution of occupational shares been known.

Firm dynamics and forecasting: We obtain our baseline industry dynamics by taking occupational shares in 2002 and 2004 and extrapolating their change eight years forward, to 2012. The baseline forecast therefore relies only on the recent trend in industry occupational shares. We compare this with a simple leader model that additionally incorporates contemporaneous occupational shares among leading firms, and with an extended leader model that also includes their lagged occupational shares. We then evaluate the models out of sample. Using the occupational structure observed in 2013 to 2015, both for the industry as a whole and for the newly identified leaders, we apply the same estimated forecasting relationships to predict occupational shares in 2023.

In the extended machine-learning specification, out-of-sample forecasting accuracy about industry-occupation shares improves by about 7% relative to a baseline. We also used an alternative definition of leader purely based on highest productivity: while this alternative leader definition shows similar improvements within-sample, its out-of-sample performance for 2023 is substantially worse than the baseline, which is consistent with our theoretical observation that productivity alone is not a clear predictor of adoption. This highlights the benefits of our machine-learning (ML) exercise. When we aggregate the forecast to national occupational shares, forecast accuracy improves by 23% using the extended ML-leaders specification rather than our baseline time trend, while performance again deteriorates if we use productivity-leaders. The forecasting gains of the extended ML-leader model persist across a range of robustness checks: heterogeneous industry dynamics at the 1-digit level; no industry split between large and small firms; different leader shares (10% or 30% of industry employment); smaller firm sizes (more then 10 or more than 5 employees); shorter (6 year) and longer (10 year) forecasting horizon; 4-year lagged variables (rather than the 2 year lag); and estimating the industry dynamics on a different time horizon than the leader characteristics.

Our leader model outperforms the occupational composition we infer from projections published by the French public administration at the national level. This benchmark is roughly on par with our time trend baseline, though it outperforms a passive no change benchmark by 9%. Their projections combine expected changes in sectoral employment with changes in occupational structure within sectors, while our approach focuses directly on the latter.[4] We also

[4]One can in principle combine both, but we do not see further improvements from this combination.

compare our predictions to those based on feasibility of automation according to 2013 expert assessments and task overlap (Frey and Osborne, 2017; Nedelkoska and Quintini, 2018), which strongly underperform against our baseline when taken at face value and provide little or no improvement even if optimally "scaled down". Most other feasibility indices do not include a prediction component that we could take to French data, but we can correlate the predicted occupational change of our extended leader model to those indexes. Correlation is negative with the early automation indices (Frey and Osborne, 2017; Nedelkoska and Quintini, 2018), captured already by the baseline time trend. Potential labor displacement effects of recent measures of AI exposure (Felten et al., 2021), suitability for machine learning (Brynjolfsson et al., 2018), or assessments of time savings by large language models (Eloundou et al., 2024) are captured only by the leader contribution.

Prediction for 2033: We retrain our model on 2012–2022 data and forecast from 2023 to 2033. It predicts the largest gains for unskilled industrial workers (0.8pp, 9%), skilled industrial (0.6pp, 8%), skilled warehousing (0.6pp, 16%) and skilled craft workers (0.4pp, 6%); and predicts largest losses for unskilled craft workers (-1.3pp, -20%), clerks (-0.5pp, -8%), administrative executives (-0.5pp, -7%), business administrative employees (-0.4pp, -6%), surveillance officers (-0.4pp, -22%) and engineers (-0.4pp, -4%). More than half of the gains and nearly all of these losses are generated by the leaders, not by the baseline trend. The correlation with recent measures of AI exposure and large language models (Felten et al., 2021; Eloundou et al., 2024) becomes stronger and significantly negative, while interestingly the correlation with earlier automation measures (Frey and Osborne, 2017; Nedelkoska and Quintini, 2018) turns significantly positive.

Why do we machine-learn the leader characteristics and then do forecasting, rather than trying to machine-learn the occupational evolution directly? For machine learning with need many observations to train the model. Occupational change at the national level does not provide many numbers to train a model, even if we disaggregate by industry. Our approach assigns each firm a "closeness" to the future occupational structure, generating hundreds of thousands of targets to train the model. The real challenge is to validate that this really identifies useful "leaders" for forecasting, which is the challenge of our out-of-sample forecasting exercise. We apply all this to occupational change because of it societal relevance, but future work could apply a similar methodology to other features associated with technological change, such as predicting the evolution of hierarchies within the organization of the firm, the capital-labor ratios of firms, or the capital structure of firms.

We review additional related literature next, before we lay out a simple model that illustrates our approach and motivates the functional forms we use in estimation. We then describe the data and an overview of occupational change over time, characterize leading firms and assess out-of-sample forcasting performance. We assess robustness, and how this relates to measures of technological feasibility, before we apply it to predict upcoming technological change.

Additional Related literature. Beyond the literature we have already covered, this paper relates to three strands of literature. The first studies the impact of technology on labor demand. A large body of work documents how automation, computerization, and offshoring reshape the occupational distribution, through skill-biased technological change (Acemoglu and Autor, 2011), job polarization (Autor and Dorn, 2013; Goos et al., 2014), and robotization (Acemoglu and Restrepo, 2020). The task-based framework of Autor et al. (2003) provides a canonical mapping from technological change to occupational tasks in order to assess labor demand, whether through expert assessment (Frey and Osborne, 2017), human or large language model assessment (Eloundou et al., 2024), task-level survey data (Nedelkoska and Quintini, 2018), individual-level task measures that account for within-occupation heterogeneity (Arntz et al., 2016), or patent-based measures of exposure to automation and artificial intelligence (e.g. Webb, 2020). More general patent-based approaches measure exposure to technological change without restricting attention to particular technologies (Kogan et al., 2021). Most of this work has been about substitution: which tasks can a machine do instead of humans. Some recent patent-based approaches have been looking at text overlap more generally (Martinez and Moen-Vorum, 2024), and attempt to distinguish automating from augmenting technological change (Autor et al., 2024). While these approaches capture the technological potential to transform occupational tasks, they do not directly measure the extent to which firms actually implement these technologies. These papers in general do not provide a clear forecasting metric, either, because they are not explicit about the time frame in which the changes will occur or about how exactly their measure maps into labor demand. Our method is agnostic about the source of change: it captures the net effect of all forces reshaping occupational structures, as revealed by the behavior of leading firms, and attempts to map this into a clear forecasting metric.

The second strand studies technology diffusion and adoption lags across firms, going back to Griliches (1957); Mansfield (1961); Rogers (1962). A growing literature documents persistent productivity gaps between frontier and laggard firms within narrowly defined industries (Andrews et al., 2016), consistent with slow diffusion of best practices. Complementary investments in organizational capital and workplace reorganization condition how well firms absorb new technologies (Bresnahan et al., 2002; Brynjolfsson et al., 2021), and the returns to adoption of information technology itself depend on the scale of their application (Lashkari et al., 2024). We build on this evidence by showing that leading firms' occupational structures contain exploitable information about the direction of diffusion, not only its speed, and by incorporating a fixed cost nature of adoption intended to capture the advantage for firms that intend to operate at larger scale.

The third strand examines the role of firm heterogeneity in shaping aggregate labor market outcomes. The rise of superstar firms and the associated decline in the aggregate labor share (Autor et al., 2020), and the role of the between-firm component in driving trends in overall wage in-

equality (Song et al., 2019), are well documented. Hubmer and Restrepo (2021) show that the joint dynamics of firm size and factor shares rationalize aggregate trends once firm heterogeneity is taken into account, and Patel (2024) provides a firm-based account of job polarization in French data. Event study evidence further shows that technology adoption episodes trigger sizable within firm occupational reallocation (Humlum, 2022). Our contribution is to exploit firm heterogeneity for forecasting rather than for explaining contemporaneous outcomes: the same dispersion in firm characteristics that generates cross sectional inequality also contains forward looking information about occupational reallocation.

2 Model

Here we present a simple theory of technological diffusion to illuminate our empirical approach and its functional form assumptions.

Time is discrete, $t \in \{\dots, -1, 0, 1, \dots\}$. Fix an industry $i \in \mathcal{I}$ with a continuum $\mathcal{F}_{i,t}$ of firms and a stationary distribution $G_i(z, I)$ of productivity z and innovativeness I. Innovativeness is not directly observed, but other characteristics, such as age, are. Let $\bar{G}_i$ denote the joint distribution of z and I and of other observables $s = (s_1, \dots, s_N) \in \mathbb{S}$. We assume that productivity z and innovativeness I are statistically dependent on observables s, so that observables are informative about both firm characteristics.[5] Firm characteristics follow a Markov process, determine exogenous exit, and are drawn by entrants so that $\bar{G}_i$ is the ergodic distribution.[6] Let $\mathcal{F}_{i,t}(s)$ be the infinite set of firms with observables s, allowing us to apply a law of large numbers.

Let $\mathbf{\Theta}$ collect all possible technologies, and $\Theta \subset \mathbf{\Theta}$ denotes the subset that a firm has mastered and combines in production. Focus on time $t_0 = 0$, where technologies $\Theta_- \subset \mathbf{\Theta}$ are widely understood and free to operate. A larger set $\Theta_+ \supset \Theta_-$ is available at organizational cost $c(t + I + \alpha A_{t-1})$, with $c(\cdot)$ decreasing and $c(T) = 0$. Time lowers adoption costs as knowledge spreads, while I measures how far ahead of their time firms are.

Let $\mathcal{A}_t(\Theta)$ denote the identities of adopters of Θ, and $A_{t-1} = |\mathcal{A}_{t-1}(\Theta_+)|$ how many there are. Prior adoption lowers costs when $\alpha > 0$ and has been a focus in some of the adoption literature, but does not alter the main results we present. For exposition we set $\alpha = 0$ and write $C(I + t \mid \Theta)$.

A firm that operates technologies in set Θ produces output according to a production function $zF(L_1, .., L_O, K_1, .., K_r|\Theta)$ that is homothetic and decreasing returns to scale in $(L_1, .., L_O, K_1, .., K_R)$, with standard INADA conditions and strict concavity. Input L_o denotes the amount

[5] Formally, $Z \not\perp S$ and $I \not\perp S$.

[6] For notational completeness, let $\phi(z', I', s_1', .., s_N'|z, I, s_1, .., s_N)$ denote Markov chain at the firm level, $\delta(z, I, s_1, .., s_N)$ denote its probability of survival, and $\phi(z, I, s_1, .., s_N|\emptyset)$ denote the distribution for the measure μ of new entrants every period. As we will see, our production function gives no reason for endogeneous exit in this simple model.

of workers in occupations $o \in \{1, 2, ..., O\}$, and is the main focus of the paper. Input K_r allows for various types of capital $r \in 1, .., r$. We assume that mastering a larger set of technologies Θ increases output for any given input bundle. Even though we suppressed it, the production function and type distribution should be indexed by the industry under consideration. Assume that the industry under consideration is small so that its factor demands do not affect factor prices r_m for different types of capital and w_o for different types of labor. We take factor prices as constant over time, returning to this assumption below.

So the maximization problem for firm with type (z, I) in this industry at time t reduces to

$$\begin{aligned} \pi(z, \Theta) &= \max_{L_1,..,L_O,K_1,..,K_r} zF(L_1, .., L_O, K_1, .., K_M|\Theta) - \sum_{o=1}^{O} w_o L_o - \sum_{m=1}^{M} r_m K_m \\ \Pi(z, I) &= \max_{\Theta \in \{\Theta_-, \Theta_+\}} \pi(z, \Theta) - C(I + t|\Theta) \end{aligned}$$

where cost $C(I + t|\Theta_-) = 0$ and $C(I + t|\Theta_+) = c(I + t)$. Let $L_o(z, \Theta)$ be optimal employment in occupation o and $L(z, \Theta)$ total employment conditional on technology.

Across workers within a set of firms $F \subset \mathcal{F}_{i,t}$, let $S_t^{o,F}$ be occupation o's employment share and $\mathcal{S}_t^F = (S_t^{1,F}, \ldots, S_t^{O,F})$ the corresponding vector. For firm f with technology Θ, this is simply $\mathcal{S}_t^f = (L_1(z, \Theta)/L(z, \Theta), .., L_N(z, \Theta)/L(z, \Theta))$. Across all firms in the industry $\mathcal{S}_t$ abbreviates $\mathcal{S}_t^{\mathcal{F}_{i,t}}$. We assume that occupational employment shares vary across some observables.[7]

Homotheticity implies that firms using technology Θ have identical occupational shares $\mathcal{S}(\Theta)$, independently of productivity.[8] Let a_t denote the share of employment in firms with new technology set Θ_+, we can write the industry dynamics as

$$\mathcal{S}_t = a_t \underbrace{\mathcal{S}_t^{\mathcal{A}_t(\theta_+)}}_{\text{occ. distr among adopters at time } t} + (1 - a_t) \underbrace{\mathcal{S}_t^{\mathcal{A}_t(\Theta_-)}}_{\text{occ. distr of laggards at time } t} \quad (1)$$

We now examine which firm rankings identify early adopters and predict occupational change:

Proposition 1. *Fix time $T' > 0$ large enough (e.g., $T' = T$). Assign to each set of observables a rank $R(s_1, ..., s_n)$ that is strictly decreasing in the geometric distance $\|\mathcal{S}_{T'} - \mathcal{S}_0^{F_{i,0}(s_1,...,s_N)}\|$. It holds that:*

(a) Firms with higher z do not necessarily have higher rates of adoption of the new technology nor a higher fraction of the workforce employed under the new technology. This holds even conditional on innovativeness I.

[7] Let $F_{i,t}(\mathbb{S}') \subset \mathcal{F}_{i,t}$ be the firms with observables in set $\mathbb{S}' \subset \mathbb{S}$. We assume that there exists disjoint $\mathbb{S}' \subset \mathbb{S}$ and $\mathbb{S}'' \subset \mathbb{S}$ such that $F_{i,t}(\mathbb{S}')$ and $F_{i,t}(\mathbb{S}'')$ both have strictly positive measure and $\mathcal{S}^{F_{i,t}(\mathbb{S}')} \neq \mathcal{S}^{F_{i,t}(\mathbb{S}'')}$. This condition is easily testable by comparing observable occupational shares across subgroups of firms with different observable characteristics. A sufficient condition for this a set of observables that correlate with I but not with z.

[8] We review this known result in Appendix A.1. In the actual estimation we initially allow for non-homotheticities but find no improvement in forecasts from this.

(b) *At time zero, we can rank firm observables by $R(\cdot)$: any set that is ranked higher has a larger fraction of workers employed under the new technology at time zero.*

(c) *Choose a binding threshold r and define leaders L_0 as the set of firms whose observables at time zero outrank r. Let $\mathcal{L}_0 = \mathcal{S}_0^{L_0}$ denote the occupational distribution across leaders at time zero.*[9] *The occupational dynamics in* (1) *can be written as*

$$\mathcal{S}_t = \beta_{L,t} \underbrace{\mathcal{L}_0}_{\text{occ. distr among early adopters at } t_0=0} + \beta_{S,t} \underbrace{\mathcal{S}_0}_{\text{overall occ. distr at } t_0=0} \tag{2}$$

where $\beta_{L,t}$ and $\beta_{S,t}$ are scalars that depend on the horizon t.

Defining leaders by productivity, as we do in an empirical robustness check, need not identify early adopters (Proposition 1(a)). High-z firms may be less innovative or, even conditional on I, have weaker adoption incentives when the new technology saves inputs at the expense of output. See Appendix A.2.

Instead, Proposition 1(b) ranks observables by how closely their firms' initial occupational shares resemble future industry shares at time T'. Since the new technology dominates at T', closer initial shares indicate more employment under the new technology. See Appendix A.3.

For Proposition 1(c), occupational shares among leaders and the entire industry at time zero are mixtures of $\mathcal{S}(\Theta_+)$ and $\mathcal{S}(\Theta_-)$, with more weight on the former among leaders. The occupational share of the industry increasingly puts weight on $\mathcal{S}(\Theta_+)$, which can be recovered through larger weight on $\mathcal{L}_0$ relative to $\mathcal{S}_0$ (Appendix A.4).

Based on this proposition, our empirical strategy first learns the observables of firms whose initial occupational shares resemble future industry shares, then estimates the coefficients governing industry dynamics in a version of equation (1). This theorem is agnostic about whether the firms leading at time zero are still alive at time t, and whether the rest of the industry catches up because laggards adopt the technology or because laggards exit and are replaced by more innovative entrants.[10]

With more than two technologies, industry dynamics are richer but the principle remains: Assume there are also intermediate technology sets Θ' and Θ'' that became available at some point

[9] Let $F_t(s_1, .., s_N)$ denote the identity of firms with observables $s_1, .., s_N$ at time t. Then $\mathcal{L}_t = \mathcal{S}^{L_t}$ where $L_t = \cup_{\{(s_1,..,s_N):R(s_1,..,s_N)\geq r\}} F_{i,t}(s_1, .., s_N)$. When we talk about h-period lagged occupational shares of leaders, we do not lag the leader definition but the occupational shares. That is, we can define $\mathcal{L}_{t,-h} = \mathcal{S}^{L_{t,-h}}$ where $L_{t,-h} = (\cup_{\{(s_1,..,s_N):R(s_1,..,s_N)\geq r\}} F_t(s_1, .., s_N)) \cap \mathcal{F}_{i,t-h}$. with some abuse of notation write L_{t-h} instead of $L_{t,-h}$ when the context makes clear that the leaders are at t and the occupational shares from $t-h$.

[10] Individual firm dynamics are governed by the Markov transition process $\phi(\cdot \mid \cdot)$ and the exit probability $\delta(\cdot)$; see Footnote 6. If entrants have high innovativeness I that subsequently declines, young firms tend to be leaders, and subsequent entrants drive industry-level catch-up. If innovativeness instead increases with firm age, older firms drive technological change. The framework thus accommodates different patterns of technological leadership and industry-level catch-up.

in time before period $-h$, and are ordered such that $\Theta_- \subset \Theta' \subset \Theta'' \subset \Theta_+$. We maintain throughout that larger technology sets have larger costs. A particularly convenient assumption is that the newer and larger technology bundles have less of a time advantage $\iota(\Theta)$ in the cost function, which now becomes $C(I+t|\Theta) = c(I+t+\iota(\Theta))$.[11] We can still obtain a definition of leaders as in Part 4 of Proposition 1, but have to replace the industry dynamics in (1) with

$$\mathcal{S}_t = \beta_{L_1,t}\mathcal{L}_0 + \beta_{L_2,t}\mathcal{L}_{-h} + \beta_{S_1,t}\mathcal{S}_0 + \beta_{S_2,t}\mathcal{S}_{-h}, \tag{3}$$

where the new terms $\mathcal{S}_{-h}$ ($\mathcal{L}_{-h}$) denotes the $h-$period lagged occupational distribution of the industry (the leaders) at time $-h$. This representation holds as long as the four observed share vectors are linearly independent (see Appendix A.5). Linear independence will maintained throughout.[12] Specification (3) is flexible, and special cases capture levels or differences.[13]

We use (3) in estimation, comparing it against a baseline $\mathcal{S}_t = \tilde{\beta}_{S_1,t}\mathcal{S}_0 + \tilde{\beta}_{S_2,t}\mathcal{S}_{-h}$, so future occupational shares are projected from current and lagged shares at the industry level. If there are only three technologies present the lagged leader shares are not necessary and can be omitted (a specification we will call "simple"). More than four technologies would require additional lags, which our sample length does not permit.

So far we have considered only "within sample" variation where we know the future industry shares $\mathcal{S}_{T'}$. To forecast out of sample, consider $t_1 > T$. Call $\hat{\Theta}_t$ the largest technology set that is widely understood and free to operate at time t. Smaller sets can be omitted. Call $\mathbf{\Theta}_t$ the collection of available technology sets at time t that is weakly larger than $\hat{\Theta}_t$. Assume that new technologies correspond one-to-one to earlier technologies in adoption incentives and relative firm size, as formalized below:

Assumption 1. *For all $t \in (t_1 - h, t_1 + T)$, there exists a bijective mapping $Q : \mathbf{\Theta} \to \mathbf{\Theta}$ such that for any z, I and any $\Theta', \Theta'' \in \mathbf{\Theta}_t$ with $\Theta' \supset \Theta''$:*

$$\frac{\pi(z,\Theta') - \pi(z,\Theta'')}{C(I+t|\Theta') - C(I+t|\Theta'')} = \frac{\pi(Q(\Theta')) - \pi(Q(\Theta''))}{C(I+t-t_1|Q(\Theta')) - C(I+t-t_1|Q(\Theta''))} \tag{1a}$$

if $C(I+t|\Theta') > C(I+t|\Theta'')$ and otherwise equality holds for the numerators and denominators separately; additionally

$$\frac{L(z,\Theta')}{L(z,\Theta'')} = \frac{L(z,Q(\Theta'))}{L(z,Q(\Theta''))} \tag{1b}$$

with additional starting condition $L(z,\hat{\Theta}_{t_1})/L(z',\hat{\Theta}_{t_1}) = L(z,Q(\hat{\Theta}_0))/L(z',Q(\hat{\Theta}_0))$.

[11] With ranking $\iota(\Theta_-) \geq T > \iota(\Theta') > \iota(\Theta'') > \iota(\Theta_+) = 0$ this specification exactly recovers the costs for our the Θ_- and $\Theta - +$, and places the costs for the other technologies in their middle.

[12] Linear independence is only assumed at time zero, and places a joint restriction on wages and the production function at time zero.

[13] Equation (2) is a special case of (3), and so is $\mathcal{S}_t - \mathcal{S}_0 = \beta_{L,t}\left(\mathcal{L}_0 - \mathcal{L}_{-h}\right) + \beta_{S,t}\left(\mathcal{S}_0 - \mathcal{S}_{-h}\right)$.

For any new technologies $\Theta^{'}$ and $\Theta^{''}$, function $Q(\cdot)$ specifies corresponding technologies with similar cost advantage (1a) and firm size implication (1b). As we discuss below, the first assumption is the more important one, and it might be appropriate when costs involve organizational and management changes, the costs of which arguably remain relatively constant over time.

Under Assumption 1, the earlier rank function $R(s_1, \ldots, s_N)$ and cutoff r can be used to forecast from t_1:

Proposition 2. *Suppose Assumption 1 holds. At time t_1, classify leaders using the same ranking $R(s_1, \ldots, s_N)$ and cutoff r as in the initial episode. Let $\mathcal{L}_{t_1}$ and $\mathcal{L}_{t_1-h}$ denote their current and lagged occupational shares. For $\tau \in (0, T)$, the occupational shares in the new episode satisfy, with two technologies,*

$$\mathcal{S}_{t_1+\tau} = \underbrace{\beta_{L,\tau}}_{\substack{\text{adoption dynamics} \\ \text{after time zero}}} \underbrace{\mathcal{L}_{t_1}}_{\substack{\text{occ. distr among} \\ \text{early adopters at } t_1}} + \underbrace{\beta_{S,\tau}}_{\substack{\text{adoption dynamics} \\ \text{after time zero}}} \underbrace{\mathcal{S}_{t_1}}_{\substack{\text{occ. distr of} \\ \text{industry at } t_1}} \tag{4}$$

and, with four technologies,

$$\mathcal{S}_{t_1+\tau} = \beta_{L_1,\tau}\mathcal{L}_{t_1} + \beta_{L_2,\tau}\mathcal{L}_{t_1-h} + \beta_{S_1,\tau}\mathcal{S}_{t_1} + \beta_{S_2,\tau}\mathcal{S}_{t_1-h}. \tag{5}$$

In both cases, the coefficients $\beta_{\cdot,\tau}$ are those identified in the initial episode at horizon τ.

Proposition 2 sums up our final step in our approach: To use the occupational distribution of the *new* leaders, but to apply the *old* industry dynamics coefficients in forecasting. Under Assumption 1a, the ranking $R(s_1, \ldots, s_N) \geq r$ selects leaders with the same distribution of (z, I) across episodes, and the z, I combinations that adopt the newer technology remain aligned across episodes.[14]

Note that this result does not hinge on the exact choice of cutoff r: a less precise r is accounted for by adjustments in the coefficients in (4) and (5). In our empirical analysis we indeed find that a wide range of cutoff gives similar forecast precision.

The proof of Proposition 2 in Appendix A.6 shows that as long as leaders exclusively contain early adopters of the newest technology, Assumption 1a alone is sufficient to fully characterize the longer-run implications of technology, i.e., under this assumption (3) or alternatively (5) hold exactly at $\tau = T$ when all hiring is driven by the now-costless frontier technology. For $\tau < T$ or with imperfect detection of early adopters among leaders, matching employment-weighted dynamics also requires similar relative firm sizes under adoption (Assumption 1b).

[14] Firms have exactly the same adoption incentives as in the previous episode, yielding identical adoption dynamics in both episodes, which also clarifies why our results also hold if the amount of previous adoption affects costs (i.e., $\alpha > 0$ so that $\alpha|\mathcal{A}_{t-1}(\Theta_+)| > 0$ in the cost function) since these amounts would also be identical in both adoption episodes.

In sum, under assumption 1 our approach recovers the full dynamics, while precise leader characterization coupled with 1a alone enables the identification of the long-run implications of frontier adoption. If both fail, leaders may still indicate the direction of occupational change, but not its exact magnitude or timing. Forecast performance is therefore an empirical matter.

This setup is illustrative. Appendix A.7 discusses alternative modeling assumptions, such as falling capital prices and investment-specific fixed costs. Our fixed-factor-price assumption also abstracts from equilibrium wage adjustments: declining demand for an occupation may lower its wage rather than employment. A full occupational-choice model would allow for partially elastic labor supply and endogenous input prices. Our partial-equilibrium results nevertheless illustrate the direction of shifts in occupational demand, even if wage responses may attenuate them.

3 Data Description and Occupational Change Over Time

3.1 Data

Data Sources. Our analysis combines matched employer employee data from the Déclarations Annuelles de Données Sociales (DADS) with firm balance sheet and income statement data from FICUS/FARE, provided by INSEE and the DGFiP. The DADS cover nearly all salaried employment in the French private sector and report annual earnings and occupations, which we aggregate to the firm year occupation level. FICUS/FARE are based on mandatory fiscal filings and contain firm level information on sales, value added, profits, capital, investment, and other balance sheet variables. We merge the two sources using the firm identifier (SIREN). [15]

Sample Construction. We first harmonize variable definitions over time within each source. We restrict attention to firms that report at least 20 employees, in either the DADS or FICUS/FARE, in at least one year over 1994–2023, the full period covered by the data. Conditioning on a single year rather than on every year avoids dropping firms that could be selected as leaders in a given cross-section but subsequently shrink, and conversely firms that grow into the sample. We then aggregate worker information to the firm–year–occupation cell by summing (i) the number of workers and (ii) total annual earnings (wage bill). In the FICUS/FARE, we harmonize accounting variables across years, drop firms missing industry information in any year, and exclude non-relevant sectors (e.g., public sector, extraterritorial activities, and home production). We

[15] We do not use stock market variables as firm characteristics because they are available for only a very small subset of firms. At the end of 2023, 653 companies were listed on Euronext Paris or Euronext Growth Paris, compared with about 257,000 firms in our panel in the same year. Stock market information would therefore provide limited and highly selected coverage of the firm population and of individual industries. In addition, listed securities are defined at the issuer or group level, whereas our unit of observation is the legal unit (SIREN).

Table 1: Main information on the sample (2002–2023)

Statistic	**Value**
Number of observations (firm–year–occupation)	25.74 mn
Number of observations (firm–year)	5.34 mn
Number of years	21
Number of distinct firms (SIREN)	479 k
Number of occupations (2-digit)	27
Number of industries (3-digit NAF)	184
Number of industries (2-digit NAF)	49

Notes: The sample spans 2002–2023. Year 2008 is excluded. The number of years reports years with available data (21). Firm–year–occupation counts occupation cells within firms; firm–year counts unique firm–year pairs. Firms are kept only if maximum DADS employment or maximum FICUS employment over 2002–2023 (excluding 2008) reaches 20.

also drop the year 2008, in which the transition from FICUS to FARE makes the firm accounts inconsistent with adjacent years. We then merge the two sources at the firm–year level, yielding a panel that links detailed labor-market outcomes by occupation to rich firm-level balance-sheet information. Table 3.1 summarizes the resulting panel. It spans 2002–2023 (21 years, 2008 excluded) and contains 478,772 distinct firms (SIREN), 5.34 million firm–year observations, and 27 two-digit occupations. Industries appear in the data at both the three-digit (184 industries) and two-digit (49 industries) NAF levels; our baseline forecasting exercise aggregates to two-digit industries. We split the panel into two parts: an *estimation sample* spanning 2002–2012 and a *testing sample* spanning 2013–2023. The testing sample begins immediately after the end of the estimation period, so that the out-of-sample evaluation uses years not employed in estimation.

Our forecasting analysis begins in 2002. Beginning with validity year 2002, INSEE profoundly redesigned the DADS statistical production chain, moving to firm-by-firm processing with validation and corrections incorporated during production. Caliendo et al. (2015) likewise begin their DADS-based analysis in 2002, citing a change in the occupational categories used to construct their hierarchy layers. The 2003 revision of the occupational nomenclature is less consequential at the level used here, as the two-digit socio-professional categories underlying our occupational groups were preserved. Our sample nevertheless spans later measurement changes: the introduction of NAF Rev. 2 in 2008, a redesign of the DADS occupational-coding procedure in 2009, and the progressive replacement of the DADS by the DSN from 2017. We bridge the 2008 industry change at the firm level: NAF Rev. 2 codes reported from 2009 onward are mapped back to NAF Rev. 1 using INSEEs correspondence table, and each firm keeps a single industry over the whole period (its most frequent code in the first years of the sample or, for later entrants, its first observed code). Because this correspondence is not one-to-one at the three-digit level, we work

with two-digit industries. The 2009 change concerns the coding procedure, not the nomenclature: the same 27 two-digit occupational categories are observed in every year, so no crosswalk is needed for occupations.

Table 2: Descriptive statistics for key variables

Variable	**Mean**	**SD**	**P25**	**P50**	**P75**
Number of workers	86	1,960	13	22	44
Mean annual wage	21,753	23,351	10,204	18,957	28,205
Firm age (years)	19	17	7	15	26
Sales	11,131	182,612	610	1,497	4,016
Value added	3,001	60,914	247	575	1,340
Gross profits	649	23,789	4	60	212
Profitability	−0.95	286.18	0.03	0.14	0.29
Exports	1,997	59,786	0	0	1
Tangible assets	6,766	363,376	63	249	805
Intangible assets	1,430	85,595	2	32	195
Investment	513	21,064	2	13	62
ACF productivity	3.6	1.7	3.0	3.7	4.6

Notes: Firm–year level statistics for the merged DADS–FICUS/FARE panel in 2015 (261,347 firm–year observations). Monetary values are in euros for wages, and in thousands of euros for other financial variables. Profitability is gross profits over sales. ACF productivity (in logs) is estimated following Ackerberg, Caves and Frazer (2015) and is available for 99 percent of firm–year observations.

Descriptive Statistics. Table 2 reports firm-level summary statistics for 2015, the base year of the testing window, when the sample contains 261,347 firm–year observations. The median firm employs 22 workers and pays an average annual wage of roughly €19,000, while mean employment is 86 workers. Median sales and value added are €1,497k and €575k, respectively. Firms hold considerably more tangible than intangible capital at the median (€249k versus €32k), while median investment is €13k. Median exports are zero. ACF productivity, available for 99 percent of firm–year observations, has a median of 3.7 log points and an interquartile range of 1.6.

Deviations from the pre-analysis plan. Our pre-registered design estimated the forecasting model on 1997–2006 and evaluated it out of sample on 2010–2019. We depart from this design in three respects. First, we begin the analysis in 2002 rather than 1997. We were made aware that starting with validity year 2002, INSEE switched to a complete redesign of its DADS production chain after multi-year preparation (INSEE, 2012), expanding its firm coverage through a system

of "expected declarations", changing the occupational coding from manual to automated, and improving its data validation and quality assurance. The change in firm coverage would have been difficult for an analysis based on leading and lagging firms, and the change in occupational coding and its validation renders occupational measures before and after 2002 less directly comparable. With the data now extending through 2023, we therefore split the remaining sample into two equal windows: an estimation period spanning 2002–2012 (2002→2004→2012) and a testing period spanning 2013–2023 (2013→2015→2023). Second, these windows span the 2008 introduction of NAF Rev. 2 industry classification, that substantially alters the three-digit classifications. We therefore conduct the analysis at the two-digit industry level that we had initially reserved as a robustness check to bridge this 2008 change, as that classifications remains largely intact and can be cross-walked. Third, the pre-analysis plan specified robustness checks for the basic forecasting model only. We apply the same checks to both the basic and extended specifications.

3.2 Evolution of the Occupational Structure

To document the evolution of the occupational structure over time, we first define the occupational structure of the economy at time t as the vector $S_t = (S_t^o)_{o \in O}$ where S_t^o denotes the share of total employment in occupation o. Changes in the occupational structure between the years t and s are measured by the difference $S_t^o - S_s^o$. To summarize changes over time, we measure the distance between occupational distributions using the L_2 norm,

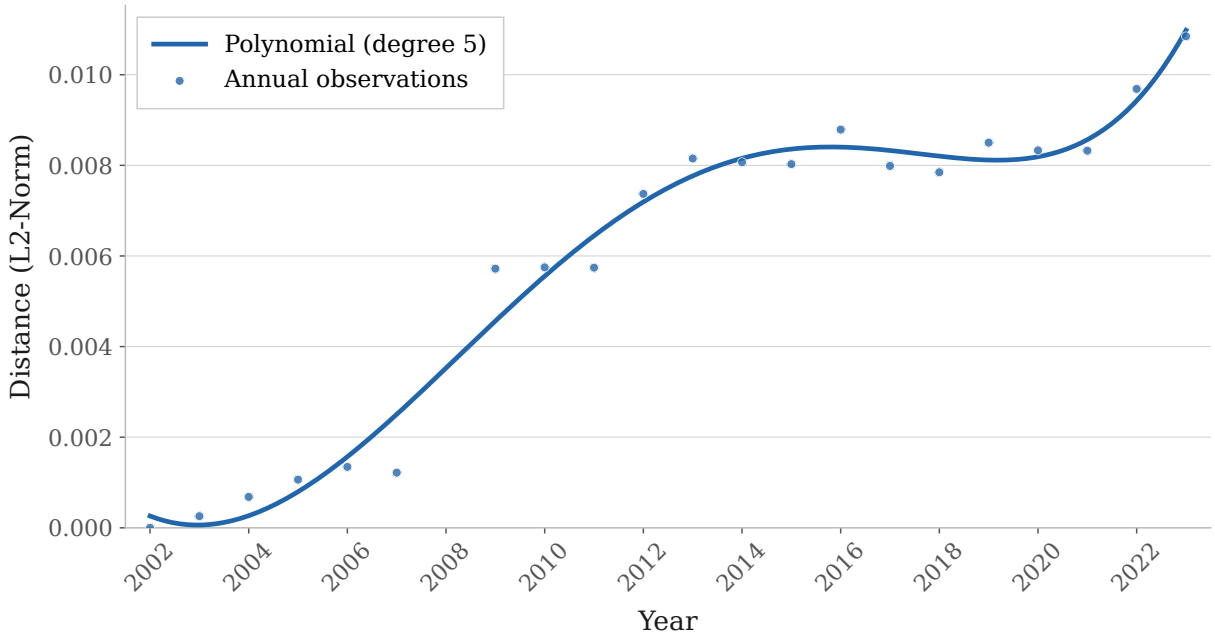
$$\|S_t - S_s\|_2 = \sqrt{\sum_{o \in O} (S_t^o - S_s^o)^2} \tag{6}$$

This metric captures the extent to which the allocation of work across occupations changes over time. Figure 1 highlights a persistent transformation of the occupational structure between 2002 and 2023. The distance of the occupational structures to the baseline year 2002 increases steadily over time, indicating that occupational change reflects a sustained structural shift rather than transitory fluctuations with mean-reversion.[16] We observe the same pattern when using the L_1 norm instead (see Figure G.1).

The underlying occupation-level changes are economically meaningful. Figure 2 shows that between 2002 and 2023, high-skill occupations, such as engineers, executives, and technicians, as well as lower-skill service occupations, expanded, while middle-skill white-collar and industrial occupations contracted. These patterns are consistent with occupational polarization. Appendix Figure G.2 presents the corresponding percentage changes and additionally highlights strong growth in health and social occupations.

[16] Because the metric is based on the L_2 norm, equally sized and same-direction year-on-year changes generate progressively larger deviations from the base year as they accumulate over time.

Figure 1: Change in the Occupational Structures over Time



Notes: The figure shows the distance between occupational structures the respective year relative to the year 2002, measured by the L_2 norm, for two-digit PCS occupations. The solid blue line is a fifth-degree polynomial fitted by ordinary least squares to the annual L_2 norm distances from the 2002 base year over 2002–2023. The time variable was rescaled as $t = (\text{year} - 2002)/21$ to improve numerical stability. The polynomial provides a smooth approximation of the observed trend. The year 2008 was excluded due to missing data.

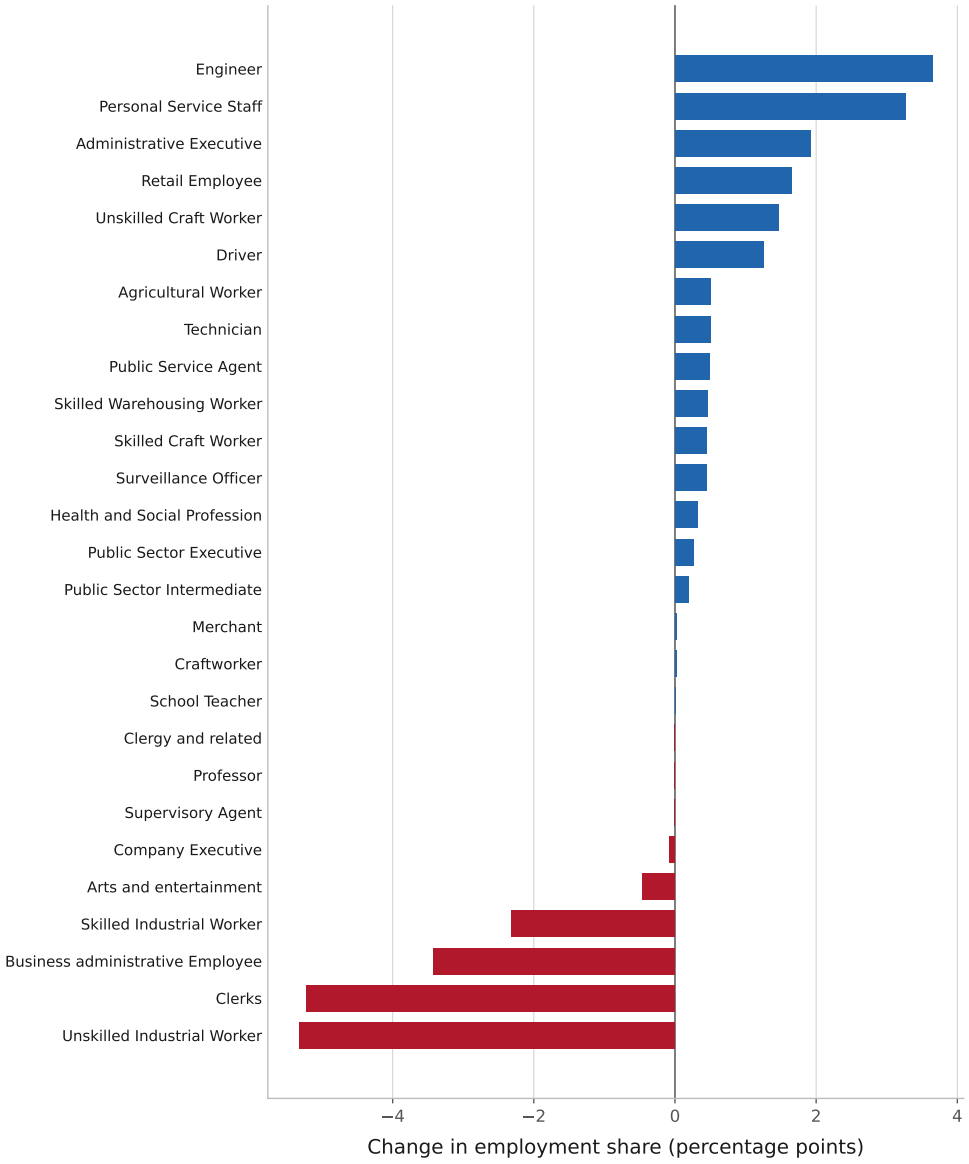
3.3 Decomposing the Occupational Change

Occupational change is driven by shifts between industries and shifts within industries and firms. We decompose changes in occupational employment to quantify the importance of each channel. The between-industry term reflects changes in industry employment shares. The within-industry term comprises occupational changes within the industry. Our theory in Section 2 focuses on the within-industry component only, but within the industry the theory is flexible: it does not matter whether existing laggards catch up or whether newborn firms adopt and laggards die out.[17] Hence, we decompose the change in the aggregate occupational employment share S_t^o between periods t and s as:

$$S_s^o - S_t^o = \underbrace{\sum_{i \in \mathcal{I}} (S_s^i - S_t^i) S_s^{o/i}}_{\text{Between industry}} + \underbrace{\sum_{i \in \mathcal{I}} (S_s^{o/i} - S_t^{o/i}) S_t^i}_{\text{Within industry}} \quad \forall o \tag{7}$$

[17] See the discussion following Proposition 1, and footnote 10.

Figure 2: Changes in the 2-digit Occupational Structures between 2002–2023



Notes: The figure shows the difference in employment shares of 2-digit PCS occupations between 2023 and 2002, in percentage points: $(S^o_{2023} - S^o_{2002}) * 100$ for all occupations o. A positive number (blue) reflects a higher share in 2023. Occupations are sorted by the size of the change. Appendix Figure G.2 reports the same changes in percent of the 2002 share.

where we denote by S^i_t the share of workers in industry i at time t, by $S^{o,i}_t$ the share of workers in occupation $o \times$ industry i at time t, and by $S^{o/i}_t$ the share of workers in occupation o within industry i at t. Hence, $S^{o/i}_t = S^{o,i}_t / S^i_t$. $\mathcal{I}$ is the set of industries.[18]

[18]The within-industry component can be separated into several sources, all of which are consistent with our basic theory: first, the reallocation of employment toward firms with different occupational mixes within the same industry due to differential firm growth; second, changes in firms' organization of production through adjustments in their

The decomposition result in Table 3 highlights three central findings. First, occupational change is overwhelmingly a within-industry phenomenon: it accounts for the majority of the change (about 98% in 2013–2022, and more than 100% in 2002–2011 when between-industry reallocation works against the aggregate shift). Second, between-industry shifts play only a limited role and, in the earlier window, even offset some of the within-industry change. Third, these patterns imply that forecasting based on sectoral growth differentials alone would miss most of the action. Instead, our results suggest that much of occupational change takes place within industries and within firms. This motivates our focus on identifying firms whose occupational structures are already tilted toward the future of their industry. Such firms are natural candidates for "leaders" in the process of occupational change. Their workforce composition may therefore contain information about the direction of future aggregate adjustments.

Table 3: Decomposition of Occupational Change

	Between industry	Within industry
2002 - 2011	-34.3%	134.3%
2013 - 2022	2.1%	97.9%

Notes: The decomposition distinguishes between-industry and within-industry contributions to occupational change averaged across all occupations. The within-industry component is the sum of between-firm, within-firm, entering-firm, exiting-firm, and net-entry components (see Appendix Table B.1 for details). Industries are defined at the two-digit NAF level; occupations at the two-digit PCS level.

4 Forecasting Occupational Change – the Role of Leading Firms

Forecasting the future of work and changes in the occupational structure of the economy is a hard task. This section lays out the structure and reports results for forecasting an industry's future occupational structure relying on "leading" firms: identifying leading firms allows to map out a catch-up process for other firms to forecast occupational change of an entire industry. We define leading firms as those whose occupational structures are ahead of broader economic trends in their industry. Our main methods for identifying leading firms use a machine learning approach and a productivity-based approach.

To validate our methodology, we first use an earlier "training" window (2002–2012) to (i) establish a method to identify leading firms and (ii) estimate the mapping from leaders occupational composition to subsequent changes in the industrys occupational structure. We then apply the

occupational composition; third, differences between the occupational structures of entering and exiting firms and the industry average; and finally, a reweighting due to differences in entry and exit shares, reflecting the effect of firm dynamics. In Appendix B, we decompose the within-industry term into these subcomponents.

estimated models to the subsequent assessment window (2013–2023), i.e., the estimated model to identify leaders based on their characteristics and the model which predicts the future occupational structure.

4.1 Variables

Our key outcome variable is the occupational structure at the industry level, defined by the vector $S_t^i = [S_t^{1,i}, \ldots, S_t^{O,i}]$, where $S_t^{o,i}$ is the share of employees in each of 27 occupational categories o at time t in industry i, defined at the two-digit NAF level (49 sectors in the baseline estimation sample). To identify leading firms, we also compute the occupational structure at the firm level, S_t^f. We measure the distance between two occupational structures S_t and S_s using the L_2 norm: $\|S_t - S_s\|_2 = \sqrt{\sum_{o \in O}(S_t^o - S_s^o)^2}$.

We construct a set of firm-level predictor variables to identify leading firms. All predictors are computed over a three-year sub-period at the start of each window: 2002–2004 for the estimation sample, and 2013–2015 for the testing sample. To avoid the Covid period we only use 2022–2023 as the sub-period for an out-of-sample prediction into the future.

We include three groups of predictors. The first consists of averages and growth rates of core firm-level variables: number of employees, total sales, value added, gross operating profits, export sales, labor cost, tangible and intangible assets, and investment. The second group captures structural characteristics of the firm through time-invariant ratios computed over the same sub-period: productivity—both labor productivity (value added per worker) and total factor productivity estimated following Ackerberg et al. (2015)—capital intensity (tangible assets relative to sales, tangible and intangible assets per worker), investment intensity (investment relative to tangible assets and per worker), profitability (gross operating surplus over sales, gross profits per worker, sales per worker), export orientation (export share of sales, exports per worker), and the average annual wage. The third group consists of firm age and two-digit industry classification, both measured at the start of each window. In total, the predictor set comprises 32 variables.

4.2 Identifying the Leading Firms

A leading firm is one whose current occupational structure resembles the future occupational structure of its industry. Intuitively, if a firm's workforce composition today looks like where its industry is heading, that firm is "ahead" of the curve. We operationalize this idea using two complementary approaches: a machine learning method that identifies leaders based on a broad set of firm characteristics, and a productivity-based method that uses total factor productivity as the sole criterion.

Machine learning approach. In the baseline specification, industries are defined at the two-digit NAF level. For each firm f in industry i, we compute the L_2-Norm between the firm's occupational structure in year t and the leave-one-out industry occupational structure h years later:

$$\left\| S_t^f - S_{t+h}^{(i_f,-f)} \right\|_2 = \sqrt{\sum_{o \in O} \left(S_t^{o,f} - S_{t+h}^{o,(i_f,-f)} \right)^2}, \tag{8}$$

where $S_{t+h}^{(i_f,-f)}$ denotes the occupational structure of all firms in industry i other than f at time $t+h$. A firm with a small distance is one whose workforce composition already anticipates where its industry is heading.

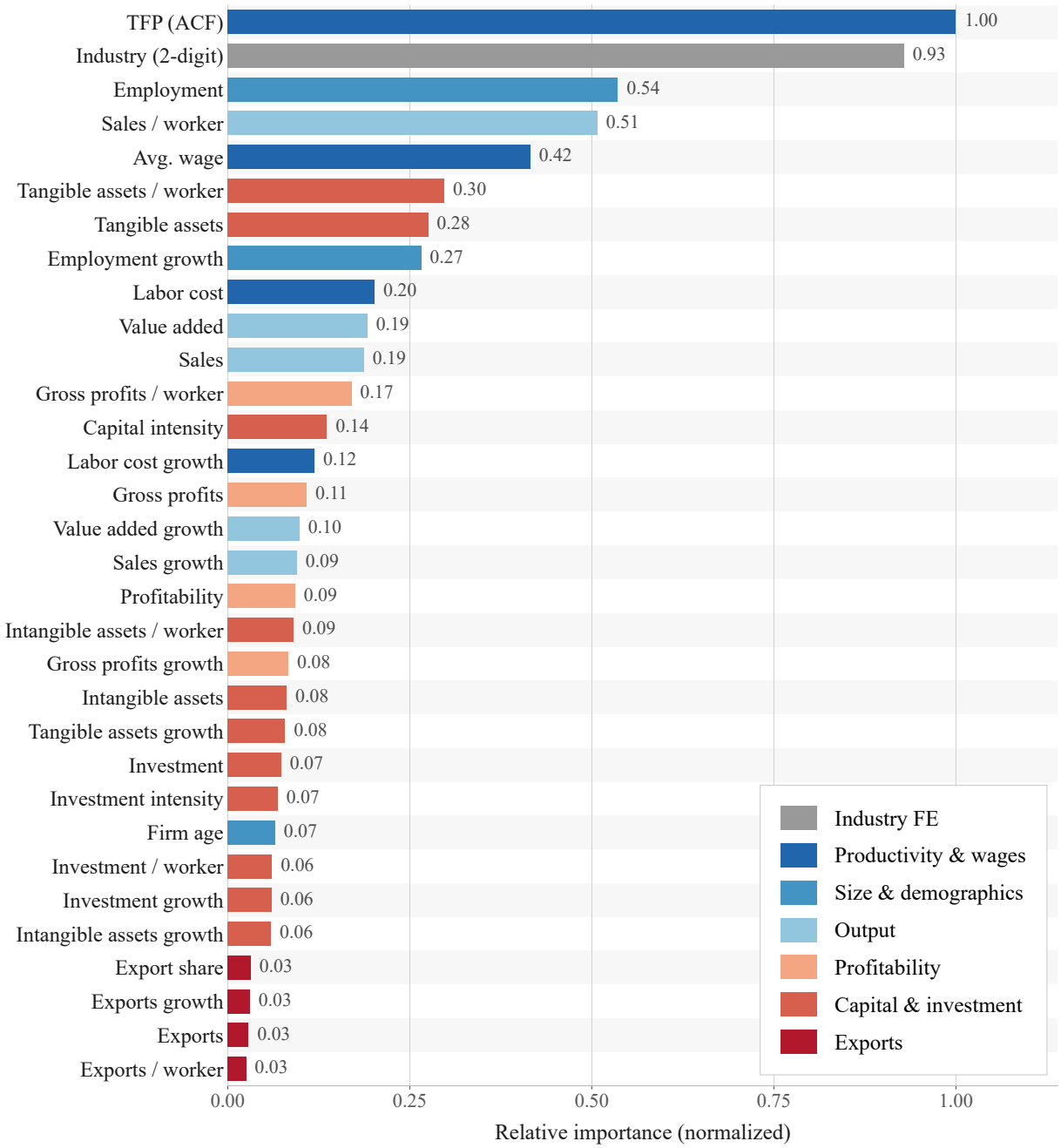
We train a random forest algorithm (Breiman, 2001) to predict this distance using the set of firm-level predictors described in Subsection 4.1. The model is trained on firms in the estimation sample, using characteristics averaged over 2002–2004 to predict occupational distances to 2012 in our baseline eight-year horizon. We tune three hyperparameters by grid search: the number of predictors considered at each split, the minimum terminal node size, and the fraction of firms sampled for each tree. The grid search selects the combination that minimizes out-of-sample prediction error on a later validation window, so that the chosen settings reward temporal generalization rather than in-sample fit. The number of trees is fixed rather than tuned, at 1,500 for the final model, well beyond the point where additional trees improve accuracy (Breiman, 2001; Probst and Boulesteix, 2018). Appendix C details the tuning protocol and the selected values. Within each industry, we rank firms by their predicted distance and classify as "ML-leaders" those with the lowest predicted distances whose cumulative employment reaches 20% of industry employment.

To account for potential differences in the determinants of leadership across the firm size distribution, we split each industry into two groups of approximately equal total employment—subject to each group containing at least 20 firms in every year and the largest firm in the smaller group employing at least 50 workers. We identify leaders separately within each size group and then pool them, so that the merged set of leaders comprises roughly 20% of total industry employment.

Figure 3 reports the variable importance measures from the random forest estimation. ACF productivity is the most important predictor, followed closely by the two-digit industry indicator and then by employment, sales per worker, and the average wage—consistent with productivity, industry context, scale, and workforce quality mattering most for predicting occupational leadership.

Productivity approach. Our secondary approach uses productivity as the sole criterion for identifying leaders. We estimate firm-level total factor productivity using the methodology of Ackerberg et al. (2015). We assume a CobbDouglas value-added production function with three

Figure 3: Variable Importance of Random Forest Model



Notes: Impurity-based variable importance from the random forest model trained on the estimation sample (2002–2004 firm characteristics predicting occupational distances in 2012), with leaders defined within two-digit NAF industry–size cells. Higher values indicate that a variable contributes more to reducing prediction error across the 1,500 trees.

inputs: white-collar employment, blue-collar employment, and capital. Labor inputs are measured by the number of employees in the respective occupational categories, while capital is

measured by tangible assets. Purchases of raw materials and other supplies, including customs duties, serve as the proxy variable for unobserved productivity in the ACF estimation procedure. We estimate separate production function parameters for the manufacturing and services sectors, allowing input elasticities to differ between groups. Production functions are estimated separately for two broad sector groups, allowing input elasticities to differ between groups.[19] Estimation is performed over rolling multi-year windows, with firm-level productivity recovered as the residual from the estimated production function. This ensures that productivity is only estimated based on past observations with respect to the current year. Within each industry-size cell, firms whose cumulative employment-weighted productivity ranks in the top 20% are classified as "productivity leaders." We restrict the sample to firms observed in at least two out of three consecutive years.

Characterizing leaders and followers. Figure 4 compares the median characteristics of leaders and followers across 1-digit industries (macro letters), for both the estimation (2004) and testing (2015) samples. ML leaders are larger than followers (about 95% in 2004 and about 20% in 2015), invest more per worker (about 100% and 50%), hold more capital per worker (about 80% and 40%), pay higher wages (about 20% and 10%, respectively), and are older. All premiums narrow between the estimation and the testing sample, but their ranking is preserved. The within-industry investment premium implies a younger capital stock, which embodies more recent production technologies (see, e.g., Greenwood et al., 1997) and is associated with new forms of labor organization (Bartel et al., 2007), providing a natural channel through which these firms may lead occupational change. Their measure of productivity is only marginally higher, and differences in profits depend on time period.

Productivity leaders dominate on TFP and profitability, earning about two to three times the profits per worker, but are not systematically more capital intensive. In the testing sample they are less capital intensive than followers and somewhat younger, and their profile shifts between 2004 and 2015. The ML method thus captures a broader and more stable dimension of leadership, while the productivity approach isolates an efficiency margin that does not generalize to other firm characteristics.

4.3 Estimation of the Forecasting Model

Once we have defined our algorithms to identify leading firms, we estimate the forecasting model of two-digit industry-level occupational structure using linear regression. These models include as covariates contemporaneous and lagged occupational structures of each industry

[19]The sector groups are defined as (1) NAF sections AF: agriculture, mining, manufacturing, electricity and gas, water and waste management, and construction and (2) all remaining sectors.

as well as the contemporaneous and lagged occupational structures of leading firms defined according to the algorithms outlined in the previous subsection. Hence, we estimate the following model:

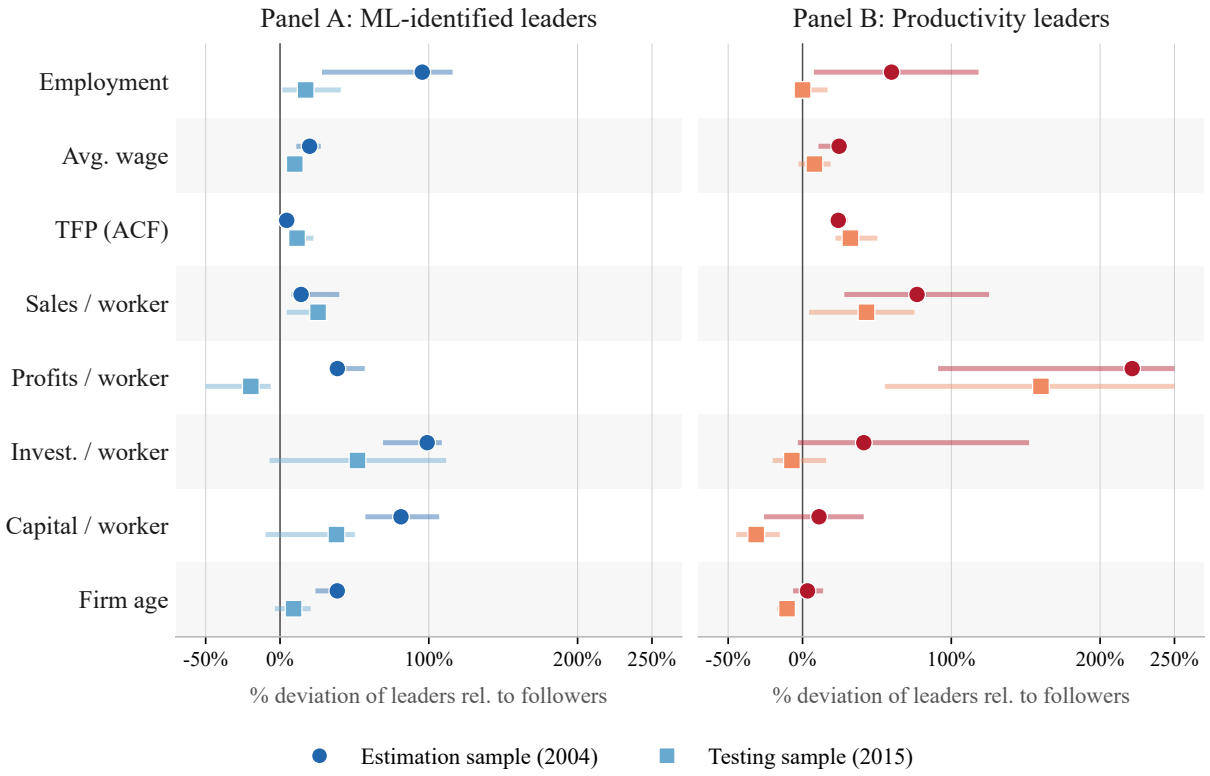
$$S_{t+k}^{o/i} = \alpha + \beta_{L_1} L_{t,t}^{o/i} + \beta_{L_2} L_{t,t-h}^{o/i} + \beta_{S_1} S_t^{o/i} + \beta_{S_2} S_{t-h}^{o/i} + \epsilon, \tag{9}$$

where $L_t^{o/i}$ is the share of workers in occupation o within leading firms in industry i at time t, while $S_t^{o/i}$ refers to the entire industry. This specification is quite flexible and includes both a specification only in contemporaneous levels or one in differences.[20] In our baseline setting we set $t = 2004$, $t - h = 2002$, and $t + k = 2012$.

The outcomes of interest are the estimates $\widehat{\alpha}$, $\widehat{\beta_{L_1}}$, $\widehat{\beta_{L_2}}$, $\widehat{\beta_{S_1}}$, and $\widehat{\beta_{S_2}}$ which we will use or prediction in the next step. We estimate different models for the productivity-leaders and the ML-

[20] Equation (9) nests the setting with only contemporaneous levels: if $\beta_{L2} = \beta_{S2} = 0$ then $S_{t+h}^{o/i} = \alpha + \beta_{L_1} L_{t,t}^{o/i} + \beta_{S_1} S_t^{o/i} + \epsilon$. Second, it also nests a model in differences: with $\beta_{L1} = -\beta_{L2}$ and $\beta_{S1} = 1 - \beta_{S2}$, we obtain $S_{t+h}^{o/i} - S_t^{o/i} = \alpha + \beta_{l_1} (L_{t,t}^{o/i} - L_{t,t-h}^{o/i}) + \beta_{s_1} (S_t^{o/i} - S_{t-h}^{o/i}) + \epsilon$.

Figure 4: Characteristics of Leaders Relative to Followers



Notes: Each dot reports the median percent deviation of leaders' characteristics relative to followers, computed across 1-digit industries. Horizontal bars span the interquartile range. Panel A shows ML-identified leaders; Panel B shows productivity leaders. Circles denote the estimation sample (2004); squares denote the testing sample (2015).

leaders. We also estimate an alternative model without including leaders which serves as our baseline model:

$$S_{t+k}^{o/i} = \tilde{\alpha} + \tilde{\beta}_{S_1} S_t^{o/i} + \tilde{\beta}_{S_2} S_{t-h}^{o/i} + \epsilon, \tag{10}$$

with the outcomes of interest being $\widehat{\tilde{\alpha}}$, $\widehat{\tilde{\beta}_{S_1}}$, and $\widehat{\tilde{\beta}_{S_2}}$.

Table 4: Forecasting Models of the Future Occupational Structure by Industry, $S_{t+h}^{o/i}$

	Baseline	Machine Learning Leaders		Productivity Leaders (ACF)	
$S_t^{o/i}$	1.3999***	1.3501***	0.3754***	0.8112***	0.8055***
	(0.0622)	(0.0614)	(0.0903)	(0.0728)	(0.0741)
$S_{t-h}^{o/i}$	−0.4539***	−0.6128***	0.3735***	−0.1179*	−0.1096*
	(0.0620)	(0.0649)	(0.0930)	(0.0632)	(0.0662)
$L_{t,t}^{o/i}$		0.1887***	1.1170***	0.2344***	0.2582***
		(0.0267)	(0.0709)	(0.0174)	(0.0590)
$L_{t,t-h}^{o/i}$			−0.9090***		−0.0258
			(0.0650)		(0.0613)
(Intercept)	0.0020***	0.0027***	0.0016***	0.0027***	0.0027***
	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Adj. R^2	0.9234	0.9261	0.9356	0.9326	0.9326
Observations	1323	1323	1323	1323	1323

Notes: OLS estimates of equations (9) and (10). The dependent variable is the share of workers in occupation o within industry i in 2012. Industries are defined at the two-digit NAF level (49 sectors; 1,323 industry–occupation cells). $S_t^{o/i}$ and $S_{t-h}^{o/i}$ are contemporaneous (2004) and lagged (2002) industry occupational structures; $L_{t,t}^{o/i}$ and $L_{t,t-h}^{o/i}$ are the corresponding structures among leading firms. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4 displays the regression output of the linear forecasting model. The extended Machine Learning model puts substantially lower weight on the industry-wide occupational structures and emphasizes the structures of the leading firms. The simple ML model leaves industry weights close to the baseline but adds a positive weight on contemporaneous leaders. Productivity leader models also emphasize the contemporaneous industry structure less than the baseline. Including leaders raises the adjusted R^2 in both setups. The sum of coefficients remains close to 1 across all five models.

4.4 Assessment of the Forecasting Models

To assess the quality of the forecasting models, we use the estimated models–precisely the estimates α and β–and generate out-of-sample forecasts for the period 2013–2023. We use the algorithms constructed in Section 4.2 to identify leading firms in 2015. We then use the models

in Equation 9 and 10 and the coefficients estimated in Section 4.3 to forecast the occupational structure of the industry in 2023. As covariates, we use the occupational structures of industries and leaders in 2013 and 2015. We assess the accuracy of these forecasts by comparing the actual occupational structure of the industry with the forecasted one using the Root Mean Square Error (RMSE). We follow the standard in the literature and compute the Theil index to compare models that incorporate leading firm information with those that do not, measuring the additional predictive value of our leader-approach:

$$\textbf{Theil ratio} = \frac{\text{RMSE (Leader Model)}}{\text{RMSE (Baseline Model)}}$$

Table 5: Assessment of the Forecasting Models

	Baseline	Machine Learning Leaders		Productivity Leaders (ACF)	
Panel A. In-Sample Assessment					
RMSE (times 10^{-2})	2.093	2.054	1.917	1.962	1.962
Theil Ratio (vs Baseline)	—	0.982	0.916	0.937	0.937
Panel B. Out-of-Sample Assessment					
RMSE (times 10^{-2})	1.302	1.236	1.216	1.604	1.592
Theil Ratio (vs Baseline)	—	0.949	0.933	1.232	1.223
Included Regressors					
$S_t^{o/i}$	✓	✓	✓	✓	✓
$S_{t-h}^{o/i}$	✓	✓	✓	✓	✓
$L_{t,t}^{o/i}$		✓	✓	✓	✓
$L_{t,t-h}^{o/i}$			✓		✓

Notes: Panel A reports in-sample results using the estimation period 2002–2012. Panel B reports out-of-sample results, applying coefficients estimated on 2002–2012 to predict 2023 occupational structures using 2013–2015 data. Industries are defined at the two-digit NAF level (49 sectors; 1,323 industry–occupation cells). RMSE is the employment weighted root mean square error across industry-occupation cells (multiplied by 10^2). The Theil ratio is the RMSE of each leader model divided by the RMSE of the baseline model; values below one indicate improved forecasting accuracy.

Table 5 reports the results for both the in-sample estimation period (2002–2012) and the out-of-sample testing period (2013–2023). We discuss each in turn. Panel (a) shows the in-sample fit. The extended ML model delivers the largest improvement, with a Theil ratio of 0.916 (roughly 8% lower RMSE than the baseline). The simple ML specification performs almost identically to the baseline (Theil ratio of 0.982), while productivity leaders yield a modest in-sample improve-

ment (Theil ratio of 0.937).

The out-of-sample results in Panel (b) are the central test. Applying the coefficients estimated on 2002–2012 to predict 2023 occupational structures using 2013–2015 data, the extended ML model achieves a Theil ratio of 0.933, a gain in forecasting accuracy of about 7% over the baseline. The simple ML model yields a ratio of 0.949, i.e., less improvement over the basline. By contrast, productivity leaders perform worse than the baseline out of sample, with Theil ratios of 1.232 and 1.223. Two features of these results stand out. First, the extended ML model retains a meaningful share of its in-sample advantage when applied to a different period and a different set of firms, indicating that the estimated relationship between leader and industry occupational structures generalizes beyond the training sample. Second, the gap between ML and productivity leaders widens sharply out of sample, consistent with the view that the broader set of firm characteristics captured by the random forest carries predictive content for occupational change that productivity alone does not.

National Occupational Forecast. For many questions the overall occupational forecast is more relevant than the industry breakdown. Table 6 aggregates the out-of-sample forecasts across industries using the 2015 industry weights, and reports the RMSE and Theil ratio for the occupational forecast at the national level. Panel A highlights that ML leaders substantially outperform our baseline, with the extended ML model achieving more than 20% improvement over the baseline. Productivity leaders again perform poorly. One reason why our ML forecast does even better at the national level might be due to noisy industry classification for firms that operate across sectors or in areas that are not neatly classified into industries. The leader-laggard distinction might still apply overall, but we might not attribute them as well to industries. Appendix Figure G.3 breaks the forecast down to each individual two-digit occupation; each time comparing the realized 2023 share with the extended ML and the baseline forecasts, with the 2015 share marked for reference.[21] The extended ML forecast is closer to the realized share than the baseline in 18 of the 27 occupations. The largest gains are for clerks (0.8 percentage points closer), business administrative employees (0.8) and engineers (0.4); the largest losses are for retail employees (0.4) and surveillance officers (0.3).

Comparison to official French forecast. To provide one comparison for our aggregate forecast, Panel B in Table 6 applies the official forecast of France Strategy to our time horizon.[22] They compare their forecast against a "naive" benchmark of constant occupational shares that are also reported in Table 6, which it outperforms, but relative to our own baseline that includes a time trend it provides hardly any improvement. The reason is that it is based mostly on predicted industry changes, and we discussed in section 3.3 that occupational change is predominantly a

[21] Appendix Figures G.4 and G.5 report the same comparison in percent of the 2015 share and isolate the leader contribution.

[22] We use the official forecasted change of France Strategy for 2013-2023 and apply 1/8th of it to the 2015 occupation shares.

Table 6: Out-of-Sample Accuracy of National Occupation-Share Forecasts, 2015–2023

Model	RMSE ($\times 10^{-3}$)	Theil ratio			
Baseline ($S_t^{o/i}$, $S_{t-h}^{o/i}$)	6.59	—			
Naive (no change in shares)	7.16	1.09			
Panel A. Leader models					
ML leaders	5.65	0.86			
ML leaders, extended	5.08	0.77			
Productivity (ACF)	8.59	1.30			
Productivity (ACF), extended	8.51	1.29			
Panel B. Official occupational projections					
France Stratégie–DARES (2015 vintage, central)	6.65	1.01			
Panel C. Technology indices					
	Direct forecast		Optimal shrinkage		
	RMSE	Theil ratio	RMSE	Theil	Shrink (%)
Computerization (Frey and Osborne, 2017)	25.88	3.93	6.40	0.97	76
Automation risk (Nedelkoska and Quintini, 2018)	10.77	1.63	7.14	1.08	88
AI exposure (Felten et al., 2021)	35.73	5.42	6.67	1.01	94
Suitability for ML (Brynjolfsson et al., 2018)	23.91	3.63	5.59	0.85	76
AI exposure (Webb, 2020)	15.82	2.40	7.16	1.09	100
GPT exposure (Eloundou et al., 2024)	12.30	1.87	6.02	0.91	65

Notes: The table reports out-of-sample RMSE (share units $\times 10^{-3}$) and Theil ratios relative to the baseline for 2023 national occupational shares. The baseline aggregates industry-level forecasts using 2015 industry weights; the naive forecast carries 2015 shares forward. Panel A reports the leader models of Table 5, evaluated on national shares. Panel B uses the France Stratégie–DARES central projection published in 2015 (*Les métiers en 2022*, 2012–2022). Projected 2023 shares are obtained by multiplying 2015 shares by $(1+g)^{8/10}$ and renormalizing, where g denotes projected ten-year growth. Panel C applies technology indices directly to 2015 occupational shares (direct forecast), or proportionally shrinks each index to minimize the Theil ratio (optimal shrinkage). The last column reports the optimal percentage reduction in each index. Since shrinkage is chosen using realized 2023 shares, these results are ex-post benchmarks rather than genuine forecasts. For Nedelkoska and Quintini (2018), the index is constructed using only French PIAAC data. The suitability-for-ML index (Brynjolfsson et al., 2018) is rescaled to $[0, 1]$ across occupations before application. Theil ratios below one indicate improved accuracy relative to the baseline.

within-sector phenomenon.

Comparison to automation technology indices. As another comparison we consider indices of automation potential. As discussed in the introduction, these indices focus on what could potentially be automated. We first ask whether our method qualitatively picks up automation potential in the sense of a negative correlation between the respective automation measure and our forecast. We then ask how we perform for actual forecasting. We consider six indices from the literature. Two early indices are based on expert assessment at the onset of the deep learning

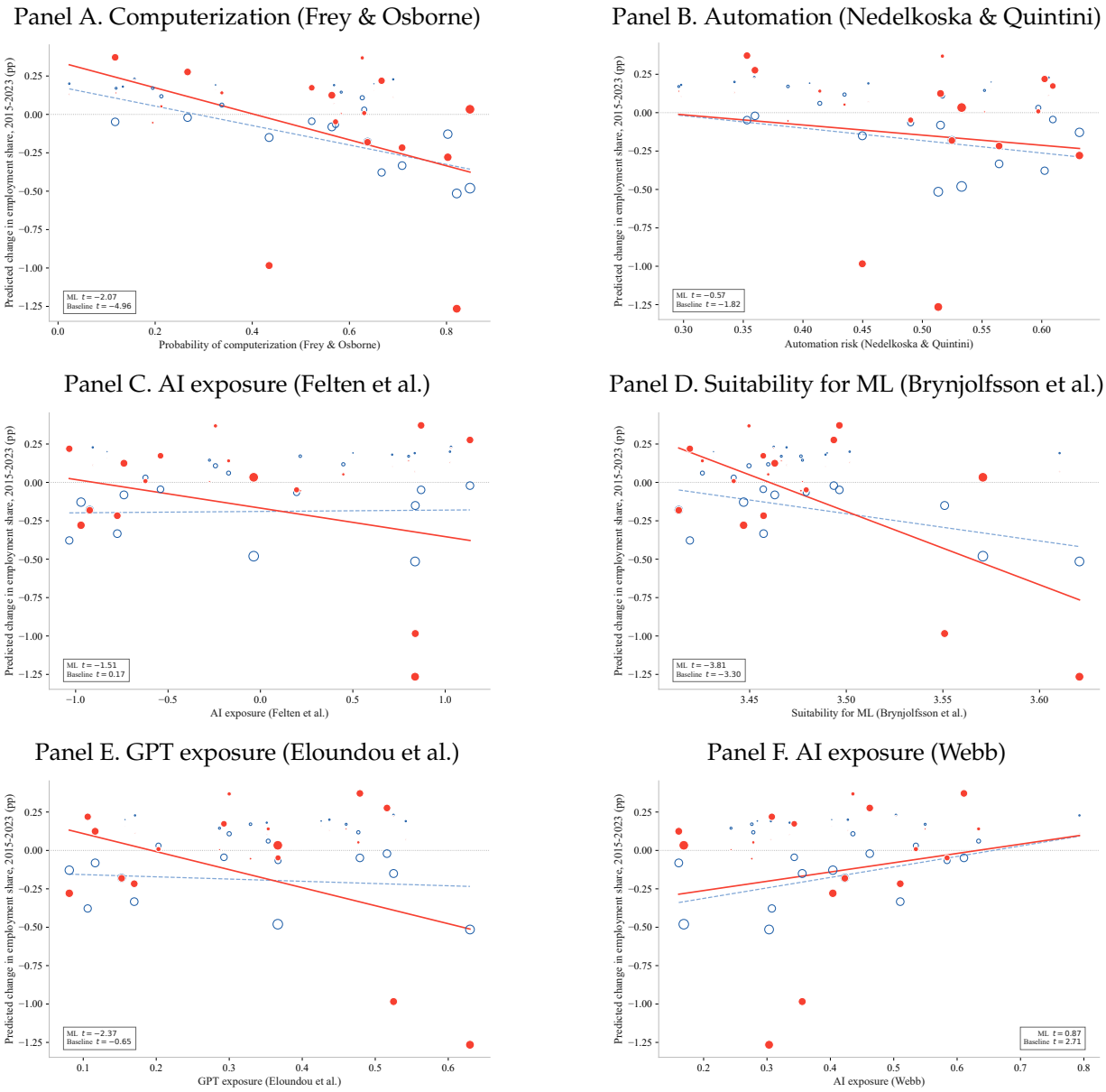
revolution, asking whether machine-computer systems can replace human labor in occupations within one to two decades. Frey and Osborne (2017) famously pioneered this and extended it to other occupations based on O*Net task overlap, while Nedelkoska and Quintini (2018) extend it via task overlap in international worker surveys, where we only use the French part. Three later measures assess the deep learning revolution as it evolved over time: either based on crowd-sourced assessments (Felten et al., 2021; Brynjolfsson et al., 2018) or on patent overlap (Webb, 2020). One measure assesses time savings of large language models (Eloundou et al., 2024) which only emerged towards the end of our forecast horizon but are correlated with earlier machine learning indices.

Figure 5 plots each index relative to forecasts of our extended ML leader (red solid) and or the baseline model (blue dotted). The two early indices by Frey and Osborne (2017) and Nedelkoska and Quintini (2018) are negatively correlated with both, the baseline model as well as the ML leader model, though not significantly for the combination of the Nedelkoska & Quintini index and the ML leader model. Hence, the baseline model already captures the risk of automation or computerization in its earlier stages with little addition by the ML leader model. In contrast, when considering the later indices in Panel C, D, and E the forecast of the ML leader model is negatively correlated with all the three indices, hence, capturing the automation risk and labor adjustments implied by these indices. Note that effects are marginally not significant for AI exposure in Panel C. The baseline model alone fails to capture the AI exposure (Felten et al., 2021) and GPT exposure (Eloundou et al., 2024). The patent-based AI exposure measure by Webb (2020) is insignificantly positively correlated with our ML leader model, plausibly because this measure is based on text similarity and, thereby, does not necessarily capture a channel of underlying substitutability but possibly also augmentation. Overall, it seems that our ML leader model is particularly suited to capture the automation potential that more recent existing indices highlight, and we will return to this discussion once we predict the occupational a decade into the future.

Given that we find some correlation with existing technology indices, one might ask if these indices do well in forecasting. We take these indices and apply them directly to the 2015 occupational shares to predict 2023 (direct forecast).[23] Alternatively, we "shrink" each index so that after its application the Theil ratio is minimized (optimal shrinkage). This allows for incomplete adjustment by 2023, but uses realized 2023 data to choose the optimal shrinkage and is therefore not a true forecast. While some indices were famously built to predict the percentage decline of occupations due to technologies that arguably fit our time-frame well (e.g. Frey and Osborne, 2017; Nedelkoska and Quintini, 2018), others were not necessarily built for this purpose and might not reflect the technologies of this time frame fully. We, therefore, do not take this exercise as a criticism of any of the indices. Rather, we see it as a cross-check whether easy application of

[23] Brynjolfsson et al. (2018) does not span $[0, 1]$, and so we first contract it into this domain.

Figure 5: Predicted Changes in Occupation Shares, 2015–2023, and Technology Indices



Notes: Each point is a two-digit PCS occupation, with marker size proportional to its 2015 employment. The horizontal axis shows the index and the vertical axis the predicted change in the occupation's share of national employment between 2015 and 2023, in percentage points, from the extended ML model (red dots, solid line) and from the baseline model without leaders (blue hollow circles, dashed line), both estimated on 2002-2004-2012 and applied out of sample from 2013-2015. Lines are weighted least squares fits, with the t-statistic of the slope reported in each panel; weights are 2015 occupation employment. Panel A uses the computerization probability of Frey and Osborne (2017), Panel B the automation risk index of Nedelkoska and Quintini (2018), Panel C the AIOE index of Felten et al. (2021), Panel D the suitability-for-machine-learning index of Brynjolfsson et al. (2018), Panel E the share of tasks exposed to large language models of Eloundou et al. (2024), and Panel F the AI exposure score of Webb (2020). Occupation 44 (Clergy) is excluded, leaving 26 occupations. Regressions of the leader contribution alone, the difference between the two predictions, on each index are reported in Appendix Table E.4.

existing feasibility indices would negate the value of our approach by already providing suitable forecasts.

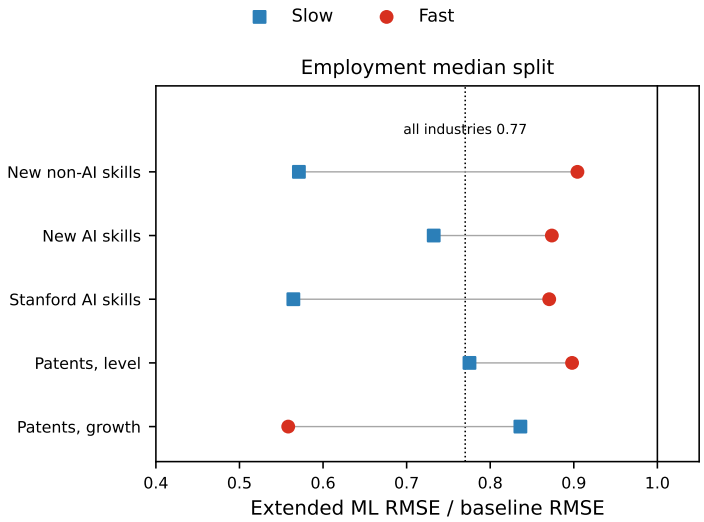
Panel C in Table 6 reports the results. All direct forecast drastically underperform the baseline as they predict too large changes. Even optimal shrinkage does not improve the implied forecasts beyond the baseline for most indices. Only one comes close with a 15% forecast improvement, but despite the shrinkage being applied in hindsight for best performance falls still short of the 23% for the extended leader model. One reason why these indices tend not to pick up occupational change is the limited overlap with actual adoption (see, e.g., Lindenlaub et al., 2026) We take this as indication that there is room for an approach like ours that focuses on adoption dynamics and is designed for forecasting.

Speed of technological progress. One might ask if our approach to forecasting only works in a relatively stable environment. Because of data availability we cannot go back to other time periods, but we can split our industries into those that are exposed more or less to technological change. We construct five measures of exposure, and each time split industries into a "fast" and a "slow" group with roughly equal employment and apply our extended ML model to forecast aggregate occupational shares within that group. Three measures are constructed from US job postings in 2010–2014 as the share of postings demanding either newly emerging non-AI (or AI) skills as in Jaumotte et al. (2026), or AI skills defined by the Stanford AI Index, all three mapped from six-digit US industries to our 49 French industries using employment weights.[24] The final two measures are based on the prevalence of automation patents following Mann and Püttmann (2023) by sector of use, once as the number of patents per year in 2010-2014 and once as the growth of that number relative to 2000-2009. See Appendix F for more details, and alternative thresholds to split industries. We see in Figure 6 that even in "fast" groups with high rates of skills change or high exposure to patents, our measure improve forecast precision by 10% to 13%, though it improves more in "slow" occupations. When the rate of patent growth changes in the last measure, our method does particularly well in fast changing environments.

Leader turnover and stability of the leader selection criteria. The theoretical model in Section 2 indicates that the machine learning approach should do well for forecasting, while productivity leaders might not capture early adoption supported by its worse than baseline out-of-sample performance. The theoretical model rests on two assumptions about the behavior of leading firms that can be assessed empirically. First, firm types (z, I) follow a Markov process, implying partial persistence in leader identity with stochastic turnover. Second, the distribution $G(z, I)$ that governs technology adoption is stationary, so that the characteristics defining leaders

[24]We rely on job postings data provided by Lightcast (2026). The emerging AI and non-AI skill definition follows the methodology in Jaumotte et al. (2026). We also use an alternative definition of AI skills based on Lightcast and Stanford Institute for Human-Centered Artificial Intelligence (2026), described in the Stanford AI Index Report 2026.

Figure 6: Extended ML error relative to the baseline error, by industry group



Notes: Ratio of the out-of-sample root mean squared error of the extended ML model to that of the baseline model for the group occupation structure. See Appendix F, Table F.1, for details. Red dots are the fast groups, blue squares the slow groups of the same ranking and split. The dotted line marks the ratio on all 49 industries together and the solid line a ratio of one.

should be stable across periods even if specific firms enter and exit the leader set. We examine whether the data are consistent with each assumption.

The Markov assumption on (z, I) implies partial persistence in leader identity: high type firms tend to remain high type, but stochastic transitions generate turnover. We find evidence for this turnover. Of the 29,071 firms classified as ML leaders in 2004, about 69% survive to 2015, a rate comparable to the overall firm survival rate of 58% (both in terms of firm counts). Among these surviving leaders, only 31% retain their leader status in 2015 (Panel B in Table G.1). Roughly two thirds of the surviving former leaders are followers in 2015. This degree of turnover is consistent with a process in which firms move in and out of the adoption frontier over an 11 year window.

The stationarity of $G(z, I)$ implies that the characteristics defining leaders should be stable across periods, even if the specific firms change. Panel A in Table G.1 shows that this is indeed the case: it reports the distance of within-industry occupational distributions between various sets of leading firms defined in 2015 and of all firms in 2015 or 2023. The distance between the leading firms of a retrained ML leader model in 2015 (*retrained 2015 leaders*) and those firms which are predicted to be leaders in 2015 when applying the ML leader model which we had estimated on 2002-2004-2012 data (*baseline rule 2015 leaders*) is smallest highlighting a persistence in the selection criteria of leaders (0.144). The distance between the set of *retrained 2015 leaders* and set of surviving ML leaders from 2004 is considerably higher (0.257). The *baseline rule 2015 leaders* are almost as close to all firms in 2023 with a distance of 0.157 as the *retrained 2015 leaders* with

0.145. This is particularly convincing because the underlying model of the *retrained 2015 leaders* was estimated to provide a close fit to the occupational distribution of all firms in 2023.

5 Robustness and Alternative Specifications

We already investigated whether our measure performs both in fast and slow-moving environments (not pre-registered). This section reports on a set of pre-registered alternative specifications and robustness checks for the forecasting exercise. Figure 7 summarizes the corresponding out-of-sample Theil ratios with *Robustness* in Panel A and *Alternative Specifications* in Panel B relative to the Theil ratio of the baseline analysis in the top row of both Panels (cf. Table 5). We start with those most closely related to our theory.

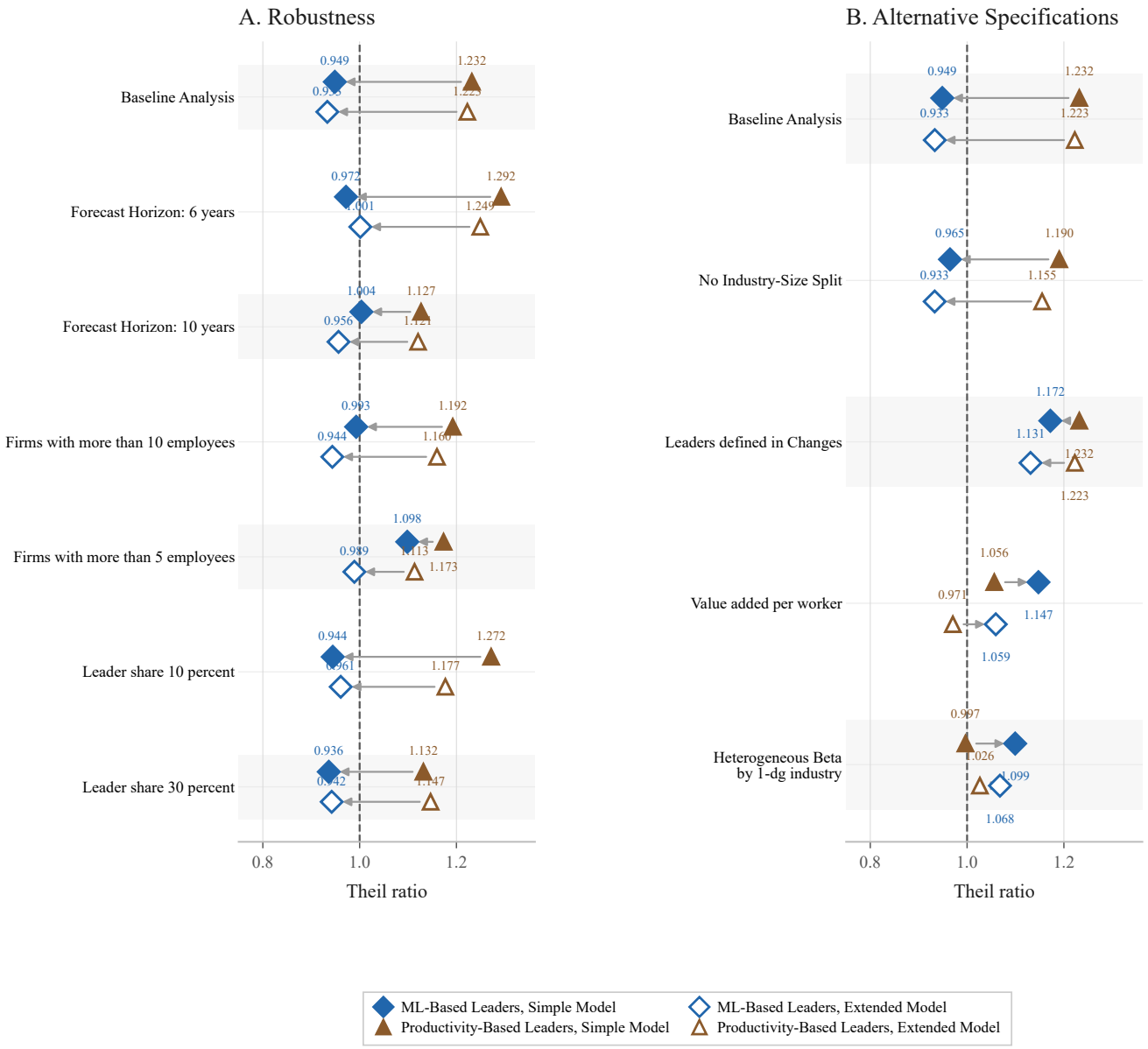
Size of leading firms. One of the more striking implications of our theory is that the threshold for the definition of leaders should not matter (Proposition 2). Our baseline sets the leader share arbitrarily at 20% of industry employment. The last two rows in Panel A in Figure 7 show that the Theil ratio for the ML leader approach is indeed unaffected when we vary the size of the leader set to 10% or 30% of industry employment. A visually even more compelling representation over more leader choices is presented in Figure G.6 in Appendix G. Leader size does matter for productivity leaders, but they do not reach an improvement over baseline.

Forecast horizon. Panel A in Figure 7 shows that with a shorter six-year forecast horizon, the ML-based simple and extended models produce Theil ratios of 0.972 and 1.001, while the productivity-based models produce 1.292 and 1.249. The simple ML model remains below one with smaller gains than in the baseline eight-year specification, while the extended ML model is essentially equal to the no-leader model. With a longer ten-year horizon, the extended ML model retains its baseline advantage with a Theil ratio of 0.956 compared to 1.004 for the simple leader model and 1.127 and 1.121 for productivity leaders.

Sample restriction on firm size. We expand the sample of firms by lowering the minimum firm-size threshold from 20 employees to ten and five employees. The extended ML model continues to outperform the no-leader model when firms with more than ten employees are included (Figure 7, Panel A). When the threshold is lowered further to five employees, the gain decreases but remains slightly below 1. These results suggest that the leader approach is not driven exclusively by the largest firms, but that occupational structures become less informative as increasingly small firms enter the sample.

Industry–size split. Our baseline definition uses two-digit NAF industries interacted with firm size. As an alternative, we remove the industry–size split and define leaders at the industry

Figure 7: Out-of-Sample Theil Ratios for Robustness & Alternative Model Specifications



Notes: The figure reports out-of-sample Theil ratios for the baseline specification and four robustness exercises (forecast horizons of six (ten) years & firms with more than 5 (10) employees) and four alternative specifications (no industry–size split, leaders defined in changes, value added per worker, and heterogeneous-β by 1-digit industry). For each specification, diamonds denote ML-based leaders and triangles denote productivity-based leaders; filled markers correspond to the simple model, hollow markers to the extended model. Thin arrows connect the productivity-based measure to the corresponding ML-based measure within each simple/extended row. The vertical dashed line at 1 indicates equality with the no-leader baseline; values below 1 indicate improved forecasting accuracy.

level only. The resulting Theil ratios shown in Panel B of Figure 7 are 0.965 and 0.933 for the ML-based simple and extended models, and 1.190 and 1.155 for the corresponding productivity-based models. Relative to the baseline, the ML gains are very comparable, while productivity leaders remain clearly worse than the no-leader benchmark. The predictive gains are therefore not driven solely by the industry–size split.

Leaders defined in changes. Our baseline specification defines leaders using levels of occupational shares. As a sensitivity check, we instead define leaders using changes in occupational structure between $t - h$ and t, aligned with the changes in the industry occupational structure between t and $t+k$. Under this definition, the ML-based simple and extended models yield Theil ratios of 1.172 and 1.131, while the productivity-based models yield 1.232 and 1.223 (Figure 7, Panel B). Defining leaders with changes eliminates the ML advantage over the baseline: both ML specifications move above one and perform not much better than the productivity approach.

Value added per worker. In Panel B of Figure 7, we present results from a specification in which we remove the ACF productivity measure from the analysis. For the productivity leaders, we replace this by value added per worker. For the ML models, we already used value added per worker along-side ACF as a predictor, and this change simply removes one of the most central variables (see Figure 3). Not surprisingly, the ML models perform substantially worse. Maybe more surprisingly, the productivity leader models perform substantially better in this specification. This contrast indicates that the performance of a univariate leader definition depends strongly on the productivity measure used. At the same time, ACF productivity appears to be particularly important for the ML models, which exploit nonlinearities and interactions with other firm characteristics to identify leading firms.

Heterogeneous-β by 1-digit industry. We also allow the forecasting coefficients (β in equation 9) to vary across 1-digit industries (Figure 7, Panel B). This additional flexibility does not improve out-of-sample performance of the ML leader models while it leads to improvements in the Theil ratios of the productivity leader specifications. Hence, the homogeneous coefficients in the baseline specification of the ML models appear to provide a useful approximation which avoids overfitting.

Additional Discussion. While this exhausts our pre-registered robustness and alternative specifications, we want to briefly touch on two points. We have seen that more parameters help within sample but can lead to overfitting out of sample. While our extended ML outperforms the simple ML in most settings, there are other ways to reduce the number of parameters in forecasting.

First, the coefficients in Table 4 sum to almost one, a restriction that the theory implies in the special case where each occupational share in the regression exactly identifies a distinct technology (see equation (1) for two technologies, and a similar logic holds for four technologies).[25] Imposing it turns the level model into a model of changes, in which the industry's future change depends on the contemporaneous gap between leaders and the industry, the gap between lagged leaders and current industry, and the industry's own recent change (Appendix D). Relative to the level baseline model, the ML-based simple and extended models in changes yield Theil ratios of 0.874 and 0.934, while the productivity-based models yield 1.133 and 1.099. The restriction also improves the no-leader model, whose Theil ratio in changes is 0.918; relative to this benchmark the simple ML model retains a 5% gain while the extended ML model does not improve on it. These specifications were not pre-registered, and we keep the level model as the baseline.

Second, occupational shares are bounded and sum to one within an industry, which a linear regressions ignores. Estimators that respect these constraints, however, make even the no-leader baseline much worse out of sample. Re-estimating equation (10) with the multivariate fractional logit of Mullahy (2015) raises the out-of-sample RMSE by 19 to 49% when shares enter in logs and more than doubles it when they enter in levels, even though the in-sample error falls. A linear model on isometric log-ratio coordinates (Aitchison, 1986) raises the out-of-sample RMSE by 23% (Appendix D). These estimators show clear indications of over-fitting in sample with poor out-of-sample properties. We therefore keep the linear model on shares, which is the most accurate benchmark we consider.

6 Predicting the Future Occupational Structure

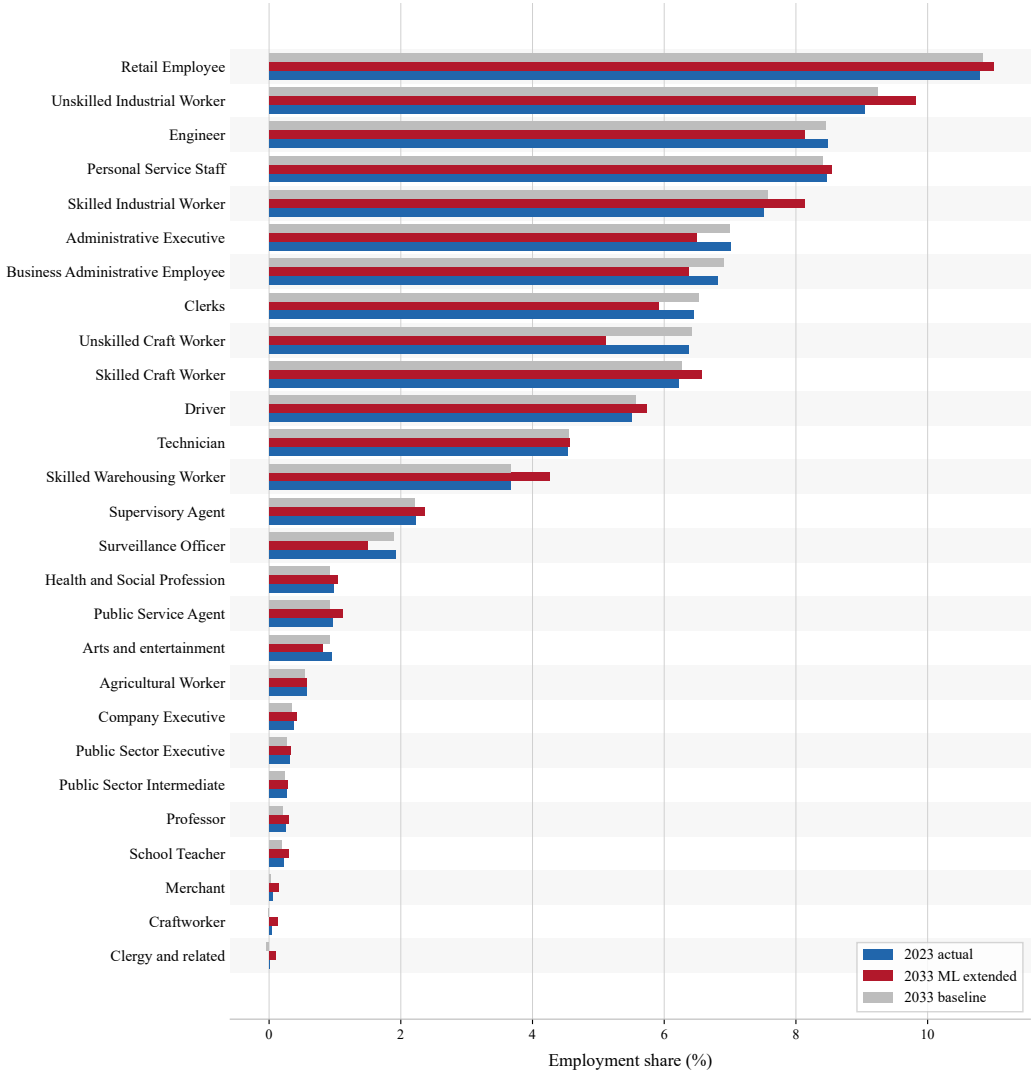
We use the extended machine-learning specification to project occupational employment shares in 2033 for the firms and sectors covered by our sample. Holding industry employment weights at their 2023 values isolates changes within industries, the principal source of occupational change documented in Table 3.

We update both leader selection and the forecasting coefficients using the most recent available ten-year outcome. The random forest is trained on firm characteristics averaged over 2010–2012 and occupational distances to 2022; the forecasting regression uses occupational structures in 2010 and 2012 to predict those in 2022. We then identify leaders from their 2021–2023 characteristics, following the selection procedure in Section 4.2, and apply the estimated relationship to industry and leader occupational structures in 2021 and 2023 to obtain the 2033 projection. The ten-year horizon extends the eight-year horizon of our baseline specification. Historical tests

[25] In a Wald test, the sum of coefficients of the extended ML model is significantly different from one at the 5%-level. Hence, we would expect difference in the prediction of the extended ML model and the respective model in changes.

also support the method at this longer horizon: the extended ML model has an out-of-sample Theil ratio of 0.956 (Figure 7).[26]

Figure 8: Predicted occupational structure in 2033 versus observed structure in 2023



Notes: Actual occupational structure in 2023 (blue) and forecast occupation structure for 2033 from the extended machine-learning model (red) and from the baseline model without leader information (light gray). The ML leader classifier and the forecasting coefficients are estimated on 2010-2012-2022 and then applied forward starting from 2021-2023. Shares aggregate industry-occupation cells to national occupation shares using 2023 industry employment shares as weights. Occupations are ordered by their 2023 share. Occupations are defined at the two-digit PCS level.

Leader information implies substantially larger changes in occupational shares than the base-

[26]The application assumes that the forecasting relationship remains informative across different cyclical conditions. The estimation window begins after the 2008–09 recession and spans the 2012–13 slowdown, whereas the application window begins during the pandemic recovery. Excluding 2020 does not eliminate possible recovery effects in the 2021–2023 observations.

line model without leaders (Figure 8). The sum of absolute share changes exceeds 7 percentage points, compared with about 1 under the baseline.[27] The difference between these forecasts measures the incremental contribution of leader information. In absolute value, this contribution exceeds 10% of the initial share in five occupations, while the baseline changes each of the fifteen largest occupations by less than 2.5% (Appendix Figures E.1 and E.2).

The projected changes favor industrial and logistics occupations, with losses concentrated in unskilled craft work and several administrative occupations. The share of unskilled industrial workers rises from 9.2% to 9.8%; skilled industrial and warehousing workers gain 0.6 percentage points each. Unskilled craft workers experience the largest decline, from 6.4% to 5.1%, followed by clerks and administrative executives, whose shares fall by 0.5 points each. The projection also departs from recent trends for some professional occupations: engineers lose 0.4 points, following a gain of 1.5 points over 2015–2023.

The gains for industrial workers do not imply a reindustrialization of employment. These occupations are defined by the type of work performed, rather than by the employer's industry. The unskilled category includes handling, sorting, packing, and shipping workers, while the skilled category includes workers in industrial maintenance and mechanical assembly. Manufacturing firms employ only 16% of unskilled and 47% of skilled industrial workers in our 2023 sample. These proportions fall to 13% and 43% in the projection, as services account for a larger share of both occupations (Appendix Table E.3). The projected gains therefore accompany a continued shift of industrial occupations toward employment in service firms.[28]

Two developments could help explain this pattern. Firms may increasingly outsource production support, maintenance, and handling to specialized service providers, shifting the recorded employer's sector while retaining industrial occupations. Growth in logistics and distribution could also raise demand for packing, sorting, and shipping workers employed by service firms. Both mechanisms are consistent with rising industrial occupational shares within services, although the projections do not distinguish their contributions.

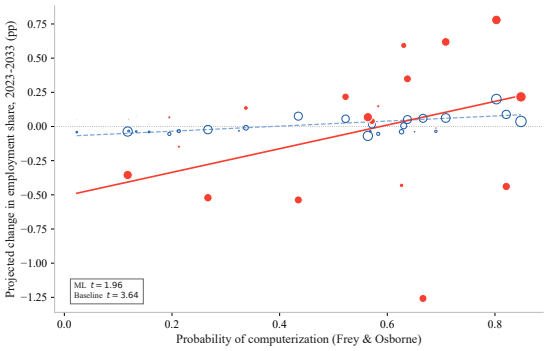
Technology exposure provides another way to characterize the projected shift in employment. As in Figure 5, we relate each automation index to the forecast of our extended ML leader (red solid) and of our baseline model (blue dotted), now focusing on forecasting into the future up to year 2033. The correlation of the two early indices by Frey and Osborne (2017) and Nedelkoska and Quintini (2018) in Panel A and B of Figure 9 with the ML leader model flips and becomes positive. The correlation of the indices with the baseline model is positive, but very small. The

[27] This statistic measures changes in occupational composition, not gross worker flows. Appendix Table E.2 reports all projected shares.

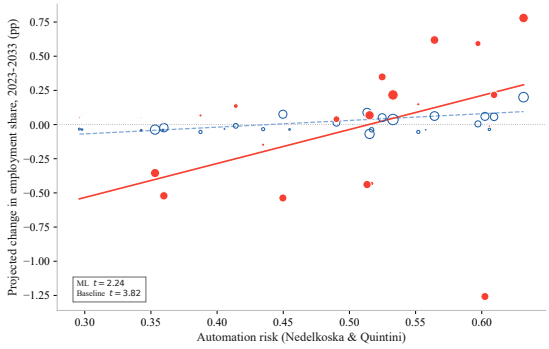
[28] Using industry weights derived from France Stratégie and DARES (France Stratégie and DARES, 2022), unskilled and skilled industrial workers gain 1.06 and 0.38 percentage points; the correlation with the fixed-weight projection is 0.97. These weights extrapolate the relative sectoral growth rates in their 2019–2030 scenario to 2023–2033.

Figure 9: Projected Changes in Occupation Shares, 2023–2033, and Technology Indices

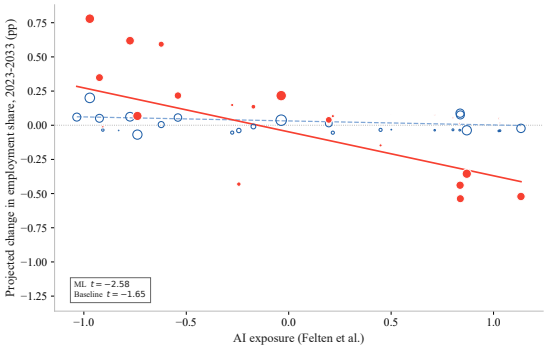
Panel A. Computerization (Frey & Osborne)



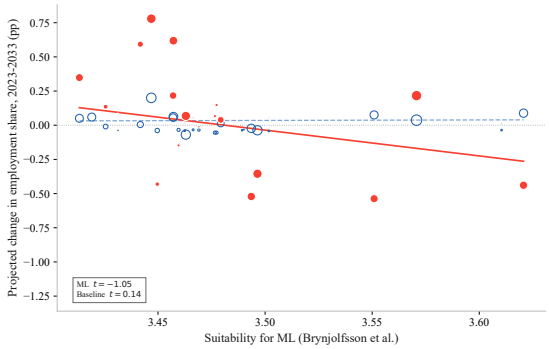
Panel B. Automation (Nedelkoska & Quintini)



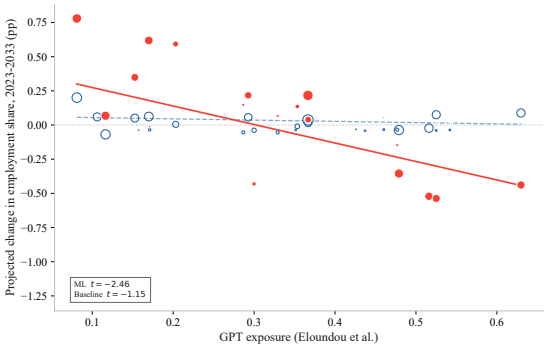
Panel C. AI exposure (Felten et al.)



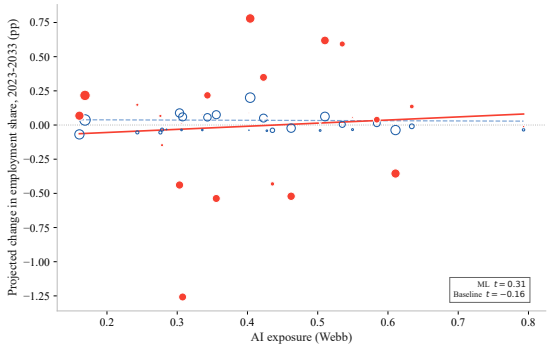
Panel D. Suitability for ML (Brynjolfsson et al.)



Panel E. GPT exposure (Eloundou et al.)



Panel F. AI exposure (Webb)



Notes: Each point is a two-digit PCS occupation, with marker size proportional to its 2023 employment. The horizontal axis shows the index and the vertical axis the projected change in the occupation's share of national employment between 2023 and 2033, in percentage points, from the extended ML model (red dots, solid line) and from the baseline model without leaders (blue hollow circles, dashed line), both estimated on 2010-2012-2022 and applied from 2021-2023. Lines are weighted least squares fits, with the t-statistic of the slope reported in each panel; weights are 2023 occupation employment. Panel A uses the computerization probability of Frey and Osborne (2017), Panel B the automation risk index of Nedelkoska and Quintini (2018), Panel C the AIOE index of Felten et al. (2021), Panel D the suitability-for-machine-learning index of Brynjolfsson et al. (2018), Panel E the share of tasks exposed to large language models of Eloundou et al. (2024), and Panel F the AI exposure score of Webb (2020). Occupation 44 (Clergy) is excluded, leaving 26 occupations. Regressions of the leader contribution alone, the difference between the two predictions, on each index are reported in Appendix Table E.4.

more recent automation indices in Panel C, D, and E are essentially uncorrelated with the baseline model. In contrast, the ML leader model is substantially negatively correlated with the recent AI- and GPT-based measures by Felten et al. (2021), Eloundou et al. (2024), and Brynjolfsson et al. (2018), though not significantly in Panel D. Hence, the leader model picks up a potential displacement effect captured in the later AI/GPT indices. The correlation with the patent-based AI exposure measure by Webb (2020) is zero, possibly because the measure captures augmentation and automation of labor. Overall, these finding are consistent with the projected employment gains for industrial and logistics occupations, which tend to score highly on the earlier automation measures and less highly on the more recent AI- and GPT-based measures.

7 Conclusion

This paper shows that forecasts of future occupational change can be improved by exploiting heterogeneity across firms and, in particular, by identifying "leading firms" whose workforce composition anticipates the future occupational structure of their industry. We develop a theoretical model building on the literature on technology diffusion and leader-laggard dynamics to show that early technology adopters are already employing the workforce of the future, thereby offering a glimpse into the future of work. Using administrative data for France, we document substantial changes in occupational structure over time, mostly arising within industries and firms. We then apply a machine-learning procedure that identifies leading firms based on a broad set of firm characteristics. Relative to a baseline without leader information, the extended machine-learning specification reduces out-of-sample forecast error at the industry-occupation level by about 7 percent and at the aggregate occupation level by about 23 percent, whereas productivity-based leaders worsen forecast accuracy. The forecasting model substantially improves accuracy even when we restrict attention to parts of the economy that is more affected by technological change. The evidence further suggests that leading firms are larger, older, better paying, more productive, and have higher sales, investment, and capital per worker than other firms in narrowly defined industries. Hence, the notion of leading firms is economically meaningful, and leadership is relatively stable in its observable characteristics over time.

A natural limitation is that the approach can only capture technologies and organizational practices that are already present in the economy, not innovations that have yet to emerge. However, conditional on the current state of technology, the model does not require expert judgment to assess which occupations will be affected or how. Instead, it uses the cross-section of firms today to extract information from those that already operate at the frontier. But since it relies on revealed adjustments in the cross section of firms rather than ex-ante judgments about specific technologies, the framework offers a practical tool for anticipating future labor-market change and for informing education, retraining, and workforce policy. In this sense the approach is

complementary to methods that forecast employment losses on the basis of expert assessment, task content analysis, or patent-based measures of technology exposure. Those approaches can, in principle, anticipate the effects of technologies that do not yet exist, but must specify which technology matters and how it maps into occupational outcomes. The leader model takes the opposite route: it is silent on the source of change but captures whatever forces are already reshaping the occupational distribution, as revealed by the firms that have moved first.

One might wonder why we did not machine-learn occupational change directly, but instead first machine learn the characteristics of leaders and only then apply it through a simple forecasting model. The answer is that we have few data points for occupational change, so machine learning is not feasible. By assigning a number to firms that characterizes their distance to the future occupational structure of the industry, we obtain hundreds of thousands of numbers and can machine-learn the characteristics of those firms that happen to be closer to the future. The real question is whether this is meaningful, or whether they happen to have an occupational structure that resembles the future by mere chance. Our out of sample forecast vindicates the approach.

While we apply the methodology to a particular dataset, future work could easily apply it to other datasets as long as they contain firm-level information on occupational structure. To the extent that leader characteristics are similar across countries, such work could also use our estimated leader characteristics to define leaders directly and then apply only the forecasting step to predict the future.

References

Acemoglu, Daron and David Autor, "Skills, Tasks and Technologies: Implications for Employment and Earnings," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 4, Amsterdam: Elsevier, 2011, pp. 1043–1171.

___ **and Pascual Restrepo**, "Robots and Jobs: Evidence from US Labor Markets," *Journal of Political Economy*, 2020, *128* (6), 2188–2244.

___ **, Gary W. Anderson, David N. Beede, Cathy Buffington, Eric E. Childress, Emin Dinlersoz, Lucia S. Foster, Nathan Goldschlag, John C. Haltiwanger, Zachary Kroff, Pascual Restrepo, and Nikolas Zolas**, "Automation and the Workforce: A Firm-Level View from the 2019 Annual Business Survey," NBER Working Paper 30659, National Bureau of Economic Research 2022.

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer, "Identification Properties of Recent Production Function Estimators," *Econometrica*, 2015, *83* (6), 2411–2451.

Aitchison, John, *The Statistical Analysis of Compositional Data*, London: Chapman and Hall, 1986.

Andrews, Dan, Chiara Criscuolo, and Peter N Gal, "The best versus the rest: the global productivity slowdown, divergence across firms and the role of public policy," 2016.

Arntz, Melanie, Terry Gregory, and Ulrich Zierahn, "The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis," OECD Social, Employment and Migration Working Papers 189, OECD Publishing, Paris 2016.

Autor, David, Caroline Chin, Anna Salomons, and Bryan Seegmiller, "New Frontiers: The Origins and Content of New Work, 1940–2018," *The Quarterly Journal of Economics*, 2024, *139* (3), 1399–1465.

___ **, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen**, "The Fall of the Labor Share and the Rise of Superstar Firms," *The Quarterly Journal of Economics*, 2020, *135* (2), 645–709.

Autor, David H. and David Dorn, "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market," *American Economic Review*, 2013, *103* (5), 1553–1597.

___ **, Frank Levy, and Richard J. Murnane**, "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, 2003, *118* (4), 1279–1333.

Bartel, Ann, Casey Ichniowski, and Kathryn Shaw, "How Does Information Technology Affect Productivity? Plant-Level Comparisons of Product Innovation, Process Improvement, and Worker Skills," *The Quarterly Journal of Economics*, 2007, *122* (4), 1721–1758.

Bick, Alexander, Adam Blandin, David Deming, Nicola Fuchs-Schündeln, and Jonas Jessen, "Measuring AI Adoption among Firms: How You Ask Matters," Federal Reserve Bank of St. Louis, On the Economy June 2026.

___ **,** ___ **, David J. Deming, and Tyler R. Schumacher**, "What Work Does Generative AI Do?," NBER Working Paper 35677, National Bureau of Economic Research 2026.

Bonney, Kathryn, Cory L. Breaux, Emin Dinlersoz, Lucia S. Foster, John C. Haltiwanger, and Aditya A. Pande, "The Microstructure of AI Diffusion: Evidence from Firms, Business Functions, and Worker Tasks," NBER Working Paper 35141, National Bureau of Economic Research 2026.

Breiman, Leo, "Random Forests," *Machine Learning*, 2001, *45* (1), 5–32.

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt, "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence," *The Quarterly Journal of Economics*, 2002, *117* (1), 339–376.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," *American Economic Journal: Macroeconomics*, 2021, *13* (1), 333–372.

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

Caliendo, Lorenzo, Ferdinando Monte, and Esteban Rossi-Hansberg, "The Anatomy of French Production Hierarchies," *Journal of Political Economy*, 2015, *123* (4), 809–852.

Comin, Diego and Bart Hobijn, "An Exploration of Technology Diffusion," *American Economic Review*, 2010, *100* (5), 2031–2059.

Doms, Mark, Timothy Dunne, and Kenneth R. Troske, "Workers, Wages, and Technology," *Quarterly Journal of Economics*, 1997, *112* (1), 253–290.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock, "GPTs are GPTs: Labor Market Impact Potential of LLMs," *Science*, 2024, *384* (6702), 1306–1308.

Felten, Edward, Manav Raj, and Robert Seamans, "Occupational, Industry, and Geographic Exposure to Artificial Intelligence: A Novel Dataset and Its Potential Uses," *Strategic Management Journal*, 2021, *42* (12), 2195–2217.

Forman, Chris, Avi Goldfarb, and Shane Greenstein, "Which Industries Use the Internet?," in "Organizing the New Industrial Economy," Vol. 12 of *Advances in Applied Microeconomics*, Emerald Group Publishing Limited, 2003, pp. 47–72.

France Stratégie and DARES, "Les métiers en 2030: Rapport du groupe Prospective des métiers et qualifications," Technical Report, France Stratégie and Dares, Paris March 2022.

Frey, Carl Benedikt and Michael A. Osborne, "The Future of Employment: How Susceptible Are Jobs to Computerisation?," *Technological Forecasting and Social Change*, 2017, *114*, 254–280.

Goos, Maarten, Alan Manning, and Anna Salomons, "Explaining Job Polarization: Routine-Biased Technological Change and Offshoring," *American Economic Review*, 2014, *104* (8), 2509–2526.

Greenwood, Jeremy, Zvi Hercowitz, and Per Krusell, "Long-Run Implications of Investment-Specific Technological Change," *The American Economic Review*, 1997, *87* (3), 342–362.

Griliches, Zvi, "Hybrid Corn: An Exploration in the Economics of Technological Change," *Econometrica*, 1957, *25* (4), 501–522.

Hampole, Menaka, Dimitris Papanikolaou, Lawrence D. W. Schmidt, and Bryan Seegmiller, "Artificial Intelligence and the Labor Market," NBER Working Paper 33509, National Bureau of Economic Research 2025.

Hubmer, Joachim and Pascual Restrepo, "Not a Typical Firm: The Joint Dynamics of Firms, Labor Shares, and Capital–Labor Substitution," Working Paper 28579, National Bureau of Economic Research March 2021. Accepted for publication in *American Economic Journal: Macroeconomics* (forthcoming); accepted-article DOI: 10.1257/mac.20230325.

Humlum, Anders, "Robot Adoption and Labor Market Dynamics," Study Paper 175, The ROCKWOOL Foundation Research Unit May 2022.

INSEE, "Les déclarations annuelles de données sociales (DADS): Guide méthodologique," Methodological guide, Institut national de la statistique et des études économiques 2012.

___ , "Séries longues sur les salaires (1950–2010) : Déclaration annuelle de données sociales (DADS) et législation sociale," June 2013. See "Historique des changements intervenus dans les déclarations annuelles de données sociales".

Jaumotte, Florence, Jaden Kim, David Koll, Elmer Li, Longji Li, Giovanni Melina, Alina Song, and Marina Mendes Tavares, "Bridging Skill Gaps for the Future," *Staff Discussion Notes*, 1 2026, *2026* (001), 1.

Koch, Michael, Ilya Manuylov, and Marcel Smolka, "Robots and Firms," *The Economic Journal*, 2021, *131* (638), 2553–2584.

Kogan, Leonid, Dimitris Papanikolaou, Lawrence D. W. Schmidt, and Bryan Seegmiller, "Technology, Vintage-Specific Human Capital, and Labor Displacement: Evidence from Linking Patents with Occupations," NBER Working Paper 29552, National Bureau of Economic Research 2021.

Lashkari, Danial, Arthur Bauer, and Jocelyn Boussard, "Information Technology and Returns to Scale," *American Economic Review*, 2024, *114* (6), 1769–1815.

Lightcast, "Lightcast global labor market dataset [Data set]," 2026. Accessed September 1, 2026.

___ **and Stanford Institute for Human-Centered Artificial Intelligence**, "The Artificial Intelligence Index Report 2026: Chapter 4 Economy – AI skills and job postings [Data set]," Stanford University 2026.

Lindenlaub, Ilse, Ryungha Oh, María Alejandra Rodríguez, and Laura Veldkamp, "Beyond Exposure: Predicting AI Adoption Based on Comparative Advantage," NBER Working Paper 35271, National Bureau of Economic Research May 2026.

Mann, Katja and Lukas Püttmann, "Benign Effects of Automation: New Evidence from Patent Texts," *The Review of Economics and Statistics*, 2023, *105* (3), 562–579.

Mansfield, Edwin, "Technical Change and the Rate of Imitation," *Econometrica*, 1961, *29* (4), 741–766.

Martinez, Joseba and Johan Moen-Vorum, "Measuring Task Exposure to Technology," 2024. Working paper.

Mullahy, John, "Multivariate Fractional Regression Estimation of Econometric Share Models," *Journal of Econometric Methods*, 2015, *4* (1), 71–100.

Nedelkoska, Ljubica and Glenda Quintini, "Automation, Skills Use and Training," Social, Employment and Migration Working Papers 202, OECD 2018.

Oshiro, Thais Mayumi, Pedro Santoro Perez, and José Augusto Baranauskas, "How Many Trees in a Random Forest?," in "Machine Learning and Data Mining in Pattern Recognition (MLDM 2012)," Vol. 7376 of *Lecture Notes in Computer Science* Springer 2012, pp. 154–168.

Patel, Aseem, "The Role of Firms in Shaping Job Polarization," Working Paper hal-04637258, HAL Open Science 2024.

Probst, Philipp and Anne-Laure Boulesteix, "To Tune or Not to Tune the Number of Trees in Random Forest," *Journal of Machine Learning Research*, 2018, *18* (181), 1–18.

Rogers, Everett M., *Diffusion of Innovations*, New York: Free Press of Glencoe, 1962.

Song, Jae, David J. Price, Fatih Guvenen, Nicholas Bloom, and Till von Wachter, "Firming Up Inequality," *The Quarterly Journal of Economics*, 2019, *134* (1), 1–50.

Webb, Michael, "The Impact of Artificial Intelligence on the Labor Market," Technical Report January 2020. Working paper, Stanford University.

A Theory: Derivations, Proofs and Discussions

A.1 Homotheticity and Occupational Shares

Homothetic functions are used to combine a notion of homogeneity with decreasing returns to scale. We review here one of their most well-known properties: conditional on technology Θ, firms choose the same input ratios independently of productivity z. Let $\mathbf{x} = (L_1, \ldots, L_O, K_1, \ldots, K_M)$ denote the vector of labor and capital inputs, with price vector $\mathbf{p}$. By homotheticity, $F(\mathbf{x} \mid \Theta) = H_\Theta(Y_\Theta(\mathbf{x}))$, where Y_Θ is homogeneous of degree one and H_Θ is increasing and concave. The Inada conditions and strict concavity of F ensure a unique interior optimum.

Conditional on Θ, the firm solves $\max_{\mathbf{x}} \{zH_\Theta(Y_\Theta(\mathbf{x})) - \mathbf{p} \cdot \mathbf{x}\}$. The first-order condition for input x_j is $zH'_\Theta(Y_\Theta)\partial Y_\Theta/\partial x_j = p_j$. Dividing by the condition for $x_1 = L_1$ gives

$$\frac{\partial Y_\Theta/\partial x_j}{\partial Y_\Theta/\partial L_1} = \frac{p_j}{w_1}. \tag{11}$$

Since Y_Θ is homogeneous of degree one, its derivatives are homogeneous of degree zero, so (11) determines $\mathbf{x}/L_1$ from prices and Θ, independently of z. Thus L_o/L_1, and hence L_o/L for $L = \sum_o L_o$, is independent of z. Using the notation from the main text, $\mathcal{S}_t^f = \left(\frac{L_1(z,\Theta)}{L(z,\Theta)}, \ldots, \frac{L_O(z,\Theta)}{L(z,\Theta)}\right) = \mathcal{S}(\Theta)$.

A.2 Proposition 1, Part (a): Productivity and Incentives to Adopt.

Let $\mathbf{x} = (L_1, \ldots, L_O, K_1, \ldots, K_M)$ denote the bundle of labor and capital inputs used in production. Consider two technologies, Θ_H and Θ_L, with Θ_H the more advanced technology. Holding I fixed, denote the net gain from adopting Θ_H by $\Delta(z, I) = \pi(z, \Theta_H) - \pi(z, \Theta_L) - [c(I + t \mid \Theta_H) - c(I + t \mid \Theta_L)]$. Since adoption costs do not depend on z, the envelope theorem gives

$$\frac{\partial \Delta(z, I)}{\partial z} = F(\mathbf{x}^*(z, \Theta_H), \Theta_H) - F(\mathbf{x}^*(z, \Theta_L), \Theta_L). \tag{A.1}$$

Thus, the incentive to adopt is increasing in productivity precisely when optimal output is higher under the new technology. Recall that our assumption on the production technology requires the new technology to produce more at any given input bundle, i.e. $F(\mathbf{x}, \Theta_H) > F(\mathbf{x}, \Theta_L)$. This does not guarantee the result in (A.1), since optimal input choices generally differ across technologies.

The following example establishes this case: Consider a single labor input L with wage normalized to one. Let $F(L, \Theta_L) = 2\sqrt{L}$ and $F(L, \Theta_H) = 2\sqrt{L} + L/(1 + L)^2$. Both technologies satisfy our production assumptions, and the new technology produces strictly more at every $L > 0$. Under the old technology, optimal labor is $L_L^* = z^2$; under the new technology, $L_H^*(z)$ solves

$z[1/\sqrt{L} + (1-L)/(1+L)^3] = 1$. The envelope theorem therefore gives

$$\frac{\partial \Delta(z,I)}{\partial z} = 2\sqrt{L_H^*(z)} + \frac{L_H^*(z)}{(1+L_H^*(z))^2} - 2z,$$

which is negative for $z \in [2,3]$. Thus, adoption incentives decrease with productivity over this range. Indeed, for $z = 2$ and $z = 3$, the respective gross profit gains from adoption are approximately 0.330 and 0.275. With an adoption cost of 0.30, the firm with $z = 2$ adopts whereas the otherwise identical firm with $z = 3$ does not. This proves Proposition 1(a).

Discussion: One can impose stronger conditions to ensure that higher productivity generates stronger incentives for adoption, but these might not necessarily capture input-saving technological progress. For example, a Hicks-neutral technological improvement immediately implies that optimal output, and hence adoption incentives, are increasing in z.[29] More generally, it is sufficient that the new technology lowers the marginal cost of output after the firm optimally adjusts its input bundle. Let

$$C(y,\Theta) = \min_{\mathbf{x}} \{\mathbf{p}\cdot\mathbf{x} : F(\mathbf{x},\Theta) \geq y\}.$$

A sufficient condition is

$$\frac{\partial C(y,\Theta_H)}{\partial y} \leq \frac{\partial C(y,\Theta_L)}{\partial y} \qquad \text{for every relevant } y. \tag{A.3}$$

With convex cost functions, (A.3) implies that optimal output under Θ_H is weakly higher than under Θ_L, and hence, by (A.1), $\partial\Delta(z,I)/\partial z \geq 0$.[30] Condition (A.3) is stronger than requiring the new technology to produce more at every given input bundle. Moreover, it need not be natural for input-saving technological change, where the benefit of adoption may precisely be to reduce the use and marginal product of particular inputs. We therefore do not impose it.

A.3 Proposition 1(b): ranking observables

Recall that a_t denotes the fraction of employment in firms within the industry with the new technology. Let $a_t(s_1,..,s_N)$ denote the fraction of employment across firm sets $\mathcal{F}_{i,t}(s_1,..,s_N)$ at time t, and let $\bar{a}_0 = \sup_{(s_1,..,s_N)} a_0(s_1,..,s_N)$ denote the suppremum.

Assumption that T' is sufficiently large amounts to $a_{T'} \geq \bar{a}$, so there is weakly more adoption at the industry level in the future than among particular observables at time zero. This is trivially

[29] Hicks-neutral technological change scales output proportionally: $F(\mathbf{x},\Theta_H) = A_H f(\mathbf{x})$ and $F(\mathbf{x},\Theta_L) = A_L f(\mathbf{x})$, with $A_H > A_L$.

[30] This assumes that the cost-minimization problem has a solution and that $C(y,\Theta)$ is convex and differentiable in output. The result follows from comparing the optimality conditions $z = C_y(y,\Theta)$; differentiability can be relaxed using subgradients.

true for $T' = T$, because this marks the point of full adoption where $a_T = 1$. We can then re-write the distance as follows:

$$\begin{aligned} & \|\mathcal{S}_{T'} - \mathcal{S}_0^{F_{i,0}(s_1,...,s_N)}\| \\ = \; & \|a_{T'}\mathcal{S}(\Theta_+) + (1-a_{T'})\mathcal{S}(\Theta_-) - (a_0(s_1,..,s_N)\mathcal{S}(\Theta_+) + (1-a_0(s_1,..,s_N))\mathcal{S}(\Theta_-))\| \\ = \; & \|(a_{T'} - a_0(s_1,...,s_N))(\mathcal{S}(\Theta_+) - \mathcal{S}(\Theta_-))\| \\ = \; & (a_{T'} - a_0(s_1,...,s_N))\|\mathcal{S}(\Theta_+) - \mathcal{S}(\Theta_-)\|. \end{aligned}$$

So this distance is rising in $a_{T'} - a_0(s_1,...,s_N)$: Closer occupational shares means a fraction employed in the new technology that is larger. This proves Proposition 1(b).

A.4 Proposition 1(c): industry dynamics

Recall that a_t denotes the fraction of employment under the new technology at the industry-level at time t, and let α_0^L denote the fraction of employment under the new technology at the industry-level at time 0 for firms with observables such that $R(s_1,..,s_N) \geq r$. Since r is binding, no all firms are leaders, and by Proposition 1(b) it holds that $\alpha_0(L) > a_0$.

We can write the occupational shares as linear combinations of the two technology-induced occupational shares $\mathcal{S}(\Theta_+)$ and $\mathcal{S}(\Theta_-)$):

$$\begin{aligned} \mathcal{S}_t &= a_t\mathcal{S}(\Theta_+) + (1-a_t)\mathcal{S}(\Theta_-), \text{ and} \\ \beta_{L,t}\mathcal{L}_0 + \beta_{S,t}\mathcal{S}_0 &= (\beta_{L,t}a_0^L + \beta_{S,t}a_0)\mathcal{S}(\Theta_+) + (\beta_{L,t}(1-a_0^L) + \beta_{S,t}(1-a_t))\mathcal{S}(\Theta_-). \end{aligned}$$

So $\mathcal{S}_t = \beta_{L,t}\mathcal{L}_0 + \beta_{S,t}\mathcal{S}_0$ if coefficients $\beta_{L,t}$ and $\beta_{S,t}$ solve

$$a_t = \beta_{L,t}a_0^L + \beta_{S,t}a_0 \tag{12}$$

$$1 - a_t = \beta_{L,t}(1-a_0^L) + \beta_{S,t}(1-a_0). \tag{13}$$

These equations solve uniquely since $a_0^L > a_0$. The proof is an explicit version of the Basis Theorem that we will apply in the case of more than two technologies (Subsection A.5). This completes the proof of Proposition 1(c).

Note: If our leaders were exactly the set of early adopters at time zero, we obtain weights $\beta_{L,t} = a_t - a_0\frac{1-a_t}{1-a_0}$ and $\beta_{S,0} = \frac{1-a_t}{1-a_0}$, which shows the shift towards leaders as adoption a_t becomes more widespread. With imperfect detection of early adopters the expressions become more complicated but the main insight remains.

A.5 The case of four technologies: deriving equation (3).

To show that equation (3) holds, note that formally, for any t, the vector $\mathcal{S}_t$ is a linear combination of the vectors $S(\theta_+)$, $S(\theta^{''})$, $S(\theta^{'})$ and $S(\theta_-)$ which form a basis (if any of these shares coincide, there would be a smaller set of basis and one would need less lagged values to accommodate this). Each of the vectors in $\{\mathcal{L}_0, \mathcal{L}_{-h}, \mathcal{S}_0, \mathcal{S}_{-h}\}$ can also be represented as a linear combination of the basis vectors. If they are linearly independent, which is the case if they each adopt a different share of technologies, by The Basis Theorem, $\{\mathcal{L}_0, \mathcal{L}_{-h}, \mathcal{S}_0, \mathcal{S}_{-h}\}$ form an alternative basis. So they also linearly represent $\mathcal{S}_t$. This completes the proof that equation (3) holds.

A.6 Proposition 2: out-of-sample forecasting

We show that the coefficients identified from the initial adoption episode also apply to the new episode. Fix $\tau \in (0, T)$ and write $t = t_1 + \tau$. The same mixture argument as for equations (2) and (3) applies at t, but we must show that the mixtures have the same weights in both episodes.

First, Assumption 1a preserves adoption choices. For a firm of type (z, I), the net payoff difference between technologies $\Theta' \supset \Theta''$ is $\pi(z, \Theta') - \pi(z, \Theta'') - [C(I + t \mid \Theta') - C(I + t \mid \Theta'')]$. When the incremental cost is positive, the firm weakly prefers Θ' precisely when the profit-gain-to-cost-gain ratio is at least one. Assumption 1a equates this ratio to the one for $Q(\Theta')$ versus $Q(\Theta'')$ at time τ. When incremental cost is zero, the assumed equality of profit and cost differences gives the same conclusion. Since Q matches the complete feasible choice sets and comparisons, each (z, I) therefore chooses corresponding technologies at $t_1 + \tau$ and τ (assuming common tie-breaking in case of indifference). By stationarity of the joint distribution of (z, I, s), the same fractions of *firms* use each corresponding technology, both in the industry and among leaders with $R(s_1, ..., s_N) \geq r$.

The forecasting equations, however, depend on fractions of *employment*, not fractions of firms. In the two-technology case, a_t in equation (1) is the fraction of all industry workers employed by firms using the advanced technology. Likewise, a_0^L and a_0 in equations (12)–(13) are employment fractions among leaders and in the industry. With four technologies, the analogous weights are the fractions of employment in each of the four technology groups. Identical adoption decisions alone do not guarantee that these weights coincide across episodes, since firms of different types may employ different numbers of workers.

Assumption 1b provides this additional result. Write $L^+(z, I)$ for a firm's optimal employment at $t_1+\tau$ and $L^0(z, I)$ for employment under its corresponding technology at τ. Its starting condition equates relative employment across productivity types under the two free technologies; hence employment under the free technology in the new episode equals $\kappa > 0$ times employment under the corresponding free technology in the old episode, with the same κ for every z. Its

technology-ratio condition then extends this proportionality to every corresponding technology choice:

$$L^+(z,I) = \kappa L^0(z,I) \qquad \text{for every } (z,I).$$

Indeed, the ratio of employment under any technology to employment under the free technology is the same in both episodes. Thus, for any corresponding group B of firms, its share of industry employment is

$$\frac{\int_B L^+(z,I)\,d\bar{G}_i}{\int L^+(z,I)\,d\bar{G}_i} = \frac{\int_B \kappa L^0(z,I)\,d\bar{G}_i}{\int \kappa L^0(z,I)\,d\bar{G}_i} = \frac{\int_B L^0(z,I)\,d\bar{G}_i}{\int L^0(z,I)\,d\bar{G}_i}.$$

The same cancellation applies when the denominator is restricted to leaders. Therefore, at corresponding dates, industry and leader occupational shares are mixtures of their respective technology- specific occupational shares with *identical employment weights*.

To conclude, collect the employment weights on the technologies for the initial leader and industry observations into a matrix W, and the weights at horizon τ into a vector w_τ. The initial forecasting coefficients solve $W\beta_\tau = w_\tau$. Linear independence in the *initial* episode identifies β_τ uniquely. We have just shown that both W and w_τ are unchanged in the new episode (after relabeling technologies through Q). Thus the same β_τ reproduces the new episode's technology mixture weights and hence its occupational shares, establishing equations (4) and (5).

Long-run implication. If leaders at the start of each episode exclusively use the frontier technology and all firms eventually adopt it, the industry's terminal occupational shares equal the initial leaders' shares. Hence a terminal forecast with coefficient one on contemporaneous leaders and zero on the remaining terms is valid even without Assumption 1b. For intermediate horizons, or when leaders include firms using older technologies, employment-weight preservation is generally needed for the exact forecast.

A.7 Discussion

Section 2 concluded with a discussion. We expand on that discussion here.

First, instead of the assumptions in our model, one could imagine that technolology is embedded in a particular capital type - say capital type M - and the cost $c(I+t+\tau_\theta)$ is not a fixed cost but the marginal cost of operating this capital with the new technology. This might be applicable if the new capital is, e.g., robots where there is a cost to install robots but the cost depends on the familiarity with such new technology. Then our basic maximization problem reads reads in analogy to (1): $\max_{\theta\in\{\theta_0,\theta_+\}} \max_{L_1,..,L_O,K_1,..,K_r} zF(L_1,..,L_O,K_1,..,K_M|\theta) - \sum_{o=1}^{O} w_o L_o - \sum_{m=1}^{M-1} r_m K_m - (r_M + C(I+t))K_M$. In this setting firms with very high innovativeness $I > T$ face simply the cost r_M and adopt factor shares according to the relative input prices. Firms with $I < T$ initially distort their production decision, but over time they converge to the production decision

of the early adopters and when $t > T$ all firms adopt the decision of early adopters. So early full adopters are still those that best foreshadow the future of the occupational structure in the industry, and we can still learn their characteristics according to our second approach. The industry will eventually catch up with their occupational distribution, though the occupational distribution along the transition is no longer just a weighted average of the initial distribution and its lag among early adopters and the whole industry as in (3), since there are now many different levels of capital M present that each induce their own occupational shares rather than than just a small set of four occupational vectors associated with the four distinct technology choices. In such a case an attempt to estimate equation (3) with an error on the right hand side would amount to linearly projecting the actual evolution into the space spanned by the occupational distribution of early full adopters and others.

Second, one might alternatively imagine that capital M which embeds the technology is formed in a putty-clay fashion where previous investment is a stock and stuck in the old technology, while new investment can choose either technology at a fixed cost but is then again locked in for the next periods. Assume that depreciation is sufficiently high that firms invest in every period. All other inputs are fully flexible. Assume that firms combine each type of capital M with other inputs in a constant returns to scale fashion to produce intermediate inputs, and these are aggregated according to some decreasing returns to scale function into final output.[31] While we do not cover this setting formally due to the dynamic nature of the choice problem, it is evident that those firms that switch part of their capital to θ_+ choose a workforce that is a weighted average of $S(\theta_+)$ and $S(\theta_0)$, and foreshadow the future evolution of the industry better than other firms that only operate the old technology. We can again learn their type. Since the leaders no longer purely operate the new technology, the weights in (1) no longer correspond to the employment share of early adopters, but now reflect the employment share in the new technology. Still we can estimate (2) or its extended version (3) to uncover the industry dynamics.

[31] Recall that a firm that produces with one technology achieves output $zF(L_1, .., L_O, K_1, ..., K_M|\theta) = f(h(L_1, .., L_O, K_1, ..., K_M))$ where $h(.)$ is constant returns to scale and $f(.)$ is concave, which is the definition of homotheticity. In principle both $h(.)$ and $f(.)$ can depend on technology θ. Here we restrict attention to functions where only $h(.)$ depends on θ, and the output of a firm that operates both technologies is given by $f(h(L'_1, .., L'_O, K'_1, ..., K'_M|\theta') + h(L''_1, .., L''_O, K''_1, ..., K''_M|\theta''))$ where variables indexed by $'$ are combined with the capital K'_M that uses technology θ', and those indexed by $''$ refers to the inputs combined with capital K''_M that embeds the new technology.

B Detailed decomposition of the Occupational Change

As described in the main text, the decomposition separates reallocation occurring across sectors from changes that arise inside industries and firms. The aggregated shift within industries and firms can be decomposed further into five components: The between-firm term captures the reallocation of employment toward firms with different occupational mixes within the same industry and is driven by differential firm growth. The within-firm term captures changes in the firm's organization of production reflecting adjustments in its occupational composition. The entry and exit terms capture that entering and exiting firms might differ in their occupational structure from the average structure in the industry. The net-entry term summarizes a re-weighting due to different entry and exit shares, thereby capturing the effect of firm dynamics. Hence, we extend the baseline decomposition of the change in the aggregate occupational employment share S_t^o between periods t and s as given in equation 7 in the following way:

$$
\begin{aligned}
S_s^o - S_t^o = & \underbrace{\sum_{i \in \mathcal{I}} (S_s^i - S_t^i) S_s^{o/i}}_{\text{Between industry}} + \underbrace{\sum_{i \in \mathcal{I}} (S_s^{o/i} - S_t^{o/i}) S_t^i}_{\text{Within industry}} \\
= & \underbrace{\sum_{i \in \mathcal{I}} (S_s^i - S_t^i) S_s^{o/i}}_{\text{Between industry}} \\
& + \underbrace{\sum_{i \in \mathcal{I}} \left(\sum_{f \in C_i} \left(\frac{S_{t+1}^f}{S_{t+1}^i} - \frac{S_t^f}{S_t^i} \right) \frac{S_{t+1}^{o,f}}{S_{t+1}^f} \right) S_t^i}_{\text{Between firm}} + \underbrace{\sum_{i \in \mathcal{I}} \left(\sum_{f \in C_{t,s}^i} \left(\frac{S_{t+1}^{o,f}}{S_{t+1}^f} - \frac{S_t^{o,f}}{S_t^f} \right) \frac{S_t^f}{S_t^i} \right) S_t^i}_{\text{Within firm}} \\
& + \underbrace{\sum_{i \in \mathcal{I}} \left(\sum_{f \in N_{t,s}^i} \left(\frac{S_{t+1}^{o,f}}{S_{t+1}^f} - \frac{S_t^{o,i}}{S_t^i} \right) \frac{S_{t+1}^f}{S_{t+1}^i} \right) S_t^i}_{\text{Entering firms}} - \underbrace{\sum_{i \in \mathcal{I}} \left(\sum_{f \in X_{t,s}^i} \left(\frac{S_t^{o,f}}{S_t^f} - \frac{S_t^{o,i}}{S_t^i} \right) \frac{S_t^f}{S_t^i} \right) S_t^i}_{\text{Exiting firms}} \\
& + \underbrace{\sum_{i \in \mathcal{I}} \frac{S_t^{o,i}}{S_t^i} \left(\sum_{f \in N_{t,s}^i} \frac{S_{t+1}^f}{S_{t+1}^i} - \sum_{f \in X_{t,s}^i} \frac{S_t^f}{S_t^i} \right) S_t^i}_{\text{Net entry at distribution in } t}
\end{aligned}
\tag{14}
$$

where we denote by S_t^i the share of workers in industry i at time t, by $S_t^{o,i}$ the share of workers in occupation $o \times$ industry i at time t, and by $S_t^{o/i}$ the share of workers in occupation o within industry i at t. Hence, $S_t^{o/i} = S_t^{o,i}/S_t^i$. $\mathcal{I}$ is the set of industries. In addition, S_t^f is defined as the share of workers in firm f at time t, $S_t^{o,f}$ as the share of workers in occupation $o \times$ firm f at time t, and $S_t^{o/f}$ as the share of workers in occupation o within firm f at t. Hence, $S_t^{o/f} = S_t^{o,f}/S_t^f$.

Finally, $C^i_{t,s}$ is the set of firms in industry i that exist in period t and s, $X^i_{t,s}$ is the set of firms in the industry i existing in t but not in s, and $N^i_{t,s}$ is the set of firms in i not existing in t but existing in s.

The further decomposition result in Table B.1 emphasize that the composition of within-industry change differs across periods. In 2002–2011, exiting firms account for the bulk of the unweighted contribution, while between-firm reallocation enters with a negative sign. In 2013–2022, within-firm adjustments dominate (about 99% of the unweighted total), with entering firms also contributing positively.

Table B.1: Decomposition of Occupational Change

	Single Components						
	Between industry	Between firms	Within firms	Entering firms	Exiting firms	Net entry	Within industry
2002 - 2011							
$\frac{1}{N_o}\sum_o \frac{x_{o,c}}{\sum_c x_{o,c}}$	-34.3%	-63.4%	7.6%	-13.6%	180.2%	23.6%	134.3%
2013 - 2022							
$\frac{1}{N_o}\sum_o \frac{x_{o,c}}{\sum_c x_{o,c}}$	2.1%	-18.3%	98.6%	20.1%	2.3%	-4.8%	97.9%

Notes: $x_{o,c}$ with $c \in [1, ..., 6]$ reflects the six components of the decomposition: between industry, between firms, within firms, entering firms, exiting firms, and net entry. The within-industry component is the sum of the five components between firms, within firms, entering firms, exiting firms, and net entry. Industries are defined at the two-digit NAF level; occupations at the two-digit PCS level.

C Random-Forest Implementation and Hyperparameter Tuning

We implement the random forest with the `ranger` package in R. Of the four hyperparameters set before training, we tune three and fix the fourth, the number of trees. The tuned hyperparameters are `mtry` (the number of predictors the algorithm may consider at each split), `min.node.size` (the minimum number of firms in a terminal node), and `sample.fraction` (the share of firms, drawn without replacement, on which each tree is grown).

We tune by grid search over $7 \times 7 \times 5 = 245$ combinations of the three hyperparameters. Each combination is trained on one window and scored on a separate, later window, and we retain the combination with the lowest out-of-sample root mean squared error. Scoring on a later window rewards settings that generalize over time rather than settings that fit the training years. Both windows use the same two-year step structure as the main analysis and consist only of years inside the pre-analysis plan estimation window, with no post-2008 data. The tuning training window measures firm characteristics over 1997–1999 and the occupational distance target over 1999–2001, that is, $(t-h, t, t+k) = (1997, 1999, 2001)$. The validation window repeats this structure as $(2002, 2004, 2006)$ and is used only to select the hyperparameters.[32] The selected values are then stored and applied unchanged in the main analysis. For the baseline level distance target, the selected values are `mtry` $= 18$, `min.node.size` $= 75$, and `sample.fraction` $= 0.63$.

The number of trees is a computational budget choice rather than a tuning dial. Forest error declines monotonically in the number of trees and converges to a limit, so more trees cannot cause overfitting (Breiman, 2001; Probst and Boulesteix, 2018), and accuracy typically plateaus within a few hundred trees (Oshiro et al., 2012), well below our settings given the size of our sample. We use 800 trees during the grid search, enough for the ranking of hyperparameter combinations to be stable while keeping 245 refits affordable, and 1,500 trees for the final model, which is estimated only once.

We tune separately for each prediction target, the L_2 distance in levels and in changes, and each target keeps its own stored hyperparameters; the values above refer to the level target. We do not re-tune for every sample variant and robustness check: in our tests the selected combinations were very close across variants, and since tuning is performed on different years and horizons than the main analysis, the procedure targets settings that are generally good rather than exactly optimal for any one specification.

[32] The results remain stable if we use different tuning windows, such as 2002, 2004, 2012, or 2004, 2006, 2014.

D Additional Forecasting Specifications

This appendix details the two checks on the forecasting equation reported at the end of Section 5. Table D.1 reports the models estimated in changes and Table D.2 the alternative estimators; Theil ratios are relative to the linear level baseline of Table 5 unless noted.

Estimation in changes. Equation (9) leaves the four coefficients unrestricted. Our theory predicts that these shares add to unity if each occupational share on the right side of the regression identifies a distinct technology (equation (1) shows this for two technologies, and the same logic extends to four technologies). Our main ML specification has coefficients that sum roughly to 1, although a Wald test shows that the sum of coefficients in Equation (9) is significantly different from one. Hence, investigating this additional restriction appears meaningful which is given by

$$S^{o/i}_{t+k} - S^{o/i}_{t} = \alpha + b_1 \left(L^{o/i}_{t,t} - S^{o/i}_{t}\right) + b_2 \left(L^{o/i}_{t,t-h} - S^{o/i}_{t}\right) + b_3 \left(S^{o/i}_{t} - S^{o/i}_{t-h}\right) + \epsilon, \tag{15}$$

where the first term is the gap between leaders and the industry, the second is the gap between lagged leaders and contemporaneous industry, and the third the industry's own recent change. With $b_2 = 0$ this is the restricted counterpart of the simple model. The restricted no-leader model keeps only the third term, and the naive forecast $S_{t+k} = S_t$ sets all of them to zero. The restriction improves the out of sample fit of the simple ML model substantially to a Theil ratio of 0.874, while the extended ML model performs as in levels (0.934) (see Table D.1). The no-leader model in changes reaches 0.918 and the naive forecast 0.939 (Panel A of Table D.2), with the simple ML model still outperforming any other model. Theil ratios of the productivity leader models stay above one in both specifications.

Alternative estimators. Panel A of Table D.2 re-estimates the no-leader model with the multivariate fractional logit of Mullahy (2015), with shares entering as regressors in levels or in logs, with a common slope or one slope per occupation, and with or without occupation fixed effects; we also add a linear model on isometric log-ratio coordinates (Aitchison, 1986), mapped back to shares. Zeros in the log-based models are replaced by two thirds of the smallest positive share in the year; using one third, or a three-year window, changes the out-of-sample ratios by less than 0.06. The striking feature of Panel A is that the baseline itself deteriorates. In units of 10^{-2}, the out-of-sample RMSE of the linear no-leader model is 1.30 (see Table 5). Every compositional version raises the RMSE as visible from the out-of-sample Theil rations which are above one for all specifications. The estimators fit the estimation window closely – visible from in-sample Theil ratios below one – but transfer poorly to a new period, the opposite of what the forecasting exercise requires, so the linear model on shares is the appropriate benchmark. Panel B estimates the leader models on simplex coordinates and reports Theil ratios relative to the simplex no-leader model: ML leaders reduce the error by 11%, productivity leaders by 5% in the simple model.

Table D.1: Forecasting Models Estimated in Changes

	Regressors			ML Leaders		Productivity Leaders	
	Gap	Lagged gap	ΔS	In-sample	Out-of-sample	In-sample	Out-of-sample
Panel A. Level models (unrestricted), for reference							
Simple				0.982	0.949	0.937	1.232
Extended				0.916	0.933	0.937	1.223
Panel B. Models in changes, equation (15)							
Simple	✓		✓	1.014	0.874	0.973	1.133
Extended	✓	✓	✓	0.891	0.934	0.971	1.099

Notes: Theil ratios: model RMSE divided by the RMSE of the level baseline of Table 5; values below one indicate lower RMSE. In-sample: estimation period 2002–2012. Out-of-sample: coefficients estimated on 2002–2012 applied to 2013–2015 data to forecast 2023. Regressors: Gap $= L_t - S_t$; Lagged gap $= L_{t-h} - S_t$; $\Delta S = S_t - S_{t-h}$ (superscripts o/i omitted); Models in changes fit $S_{t+k} - S_t$ and are scored on S_{t+k}; RMSE is employment weighted across industry–occupation cells. The no-leader benchmarks (level baseline, naive forecast, and the no-leader model in changes) are reported in Panel A of Table D.2.

Table D.2: Alternative Estimators for the Forecasting Model

	Theil ratios	
	In-sample	Out-of-sample
Panel A. No-leader model, alternative estimators (relative to linear level baseline)		
Linear, levels, contemporaneous share only	1.021	1.018
Linear, changes	1.019	0.918
Naive, $S_{t+k} = S_t$	1.038	0.939
Fractional logit, levels, common slope, FE	1.150	2.002
Fractional logit, levels, occupation slopes, FE	0.819	3.717
Fractional logit, logs, common slope, FE	0.718	1.272
Fractional logit, logs, common slope, no FE	1.036	1.187
Fractional logit, logs, occupation slopes, FE	0.650	1.489
Fractional logit, logs, occupation slopes, no FE	0.729	1.438
Simplex (isometric log-ratio)	1.088	1.227
Panel B. Leader models on the simplex (relative to simplex no-leader model)		
ML leaders, simple	1.019	0.891
ML leaders, extended	1.014	0.888
Productivity leaders, simple	0.909	0.946
Productivity leaders, extended	0.915	0.997

Notes: Theil ratios: model RMSE divided by the RMSE of the linear level baseline of Table 5 (Panel A) or of the no-leader model estimated on simplex coordinates (Panel B); values below one indicate lower RMSE. In-sample: estimation period 2002–2012. Out-of-sample: models estimated on 2002–2012 applied to 2013–2015 data to forecast 2023. RMSE is computed on shares, employment weighted across industry–occupation cells. FE: occupation fixed effects.

E Projection to 2033

The 2033 exercise applies the forecasting method evaluated in the preceding sections to the latest observed workforce structures. The results below show how leader information changes the projected occupational mix, clarify the economic meaning of the gains for industrial workers, and relate the projected shifts to measures of technology exposure.

Estimation and application. The projection retains the forecasting method assessed in the historical exercises and updates its estimates using more recent data. Table E.1 reports the coefficients estimated from occupational structures in 2010 and 2012 and outcomes in 2022; we apply them to structures in 2021 and 2023 to project 2033. In the extended ML specification, the current leader share enters with a coefficient of -0.157 and the lagged leader share with a coefficient of 0.469, both significant at the 1% level. An occupation's share being one percentage point higher among leaders at both dates is therefore associated with a 0.31 percentage point higher predicted industry share, holding industry composition fixed and before renormalization, with the weight falling on the lagged leader structure.

Table E.1: Forecasting Coefficients Used for the 2033 Projection

	Baseline	Machine Learning Leaders		Productivity Leaders (ACF)	
		Simple	Extended	Simple	Extended
$S_t^{o/i}$	0.8150^{***}	0.8309^{***}	1.1037^{***}	0.7852^{***}	0.6856^{***}
	(0.0468)	(0.0396)	(0.0466)	(0.0486)	(0.0585)
$S_{t-h}^{o/i}$	0.1951^{***}	-0.1274^{***}	-0.4415^{***}	0.2048^{***}	0.3142^{***}
	(0.0464)	(0.0417)	(0.0507)	(0.0465)	(0.0587)
$L_{t,t}^{o/i}$		0.2758^{***}	-0.1571^{***}	0.0167^{**}	0.1429^{***}
		(0.0120)	(0.0441)	(0.0075)	(0.0423)
$L_{t,t-h}^{o/i}$			0.4692^{***}		-0.1368^{***}
			(0.0461)		(0.0451)
(Intercept)	-0.0004	0.0008^{**}	0.0010^{***}	-0.0002	-0.0002
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Adj. R^2	0.9674	0.9767	0.9784	0.9675	0.9677
Observations	1323	1323	1323	1323	1323

Notes: OLS estimates of equations (9) and (10) with $t - h = 2010$, $t = 2012$, and $t + k = 2022$. The dependent variable is the share of workers in occupation o within industry i in 2022. Conventional OLS standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Occupational changes and the contribution of leaders. Leader information changes the outlook substantially for several occupations that the baseline projects to remain close to their initial shares. Table E.2 reports observed and projected shares for occupations accounting for at least

0.5% of employment in 2023, and Figure E.1 expresses projected changes relative to those initial shares. Under the baseline, the changes are smaller than 2.5% in absolute value for each of the fifteen largest occupations.

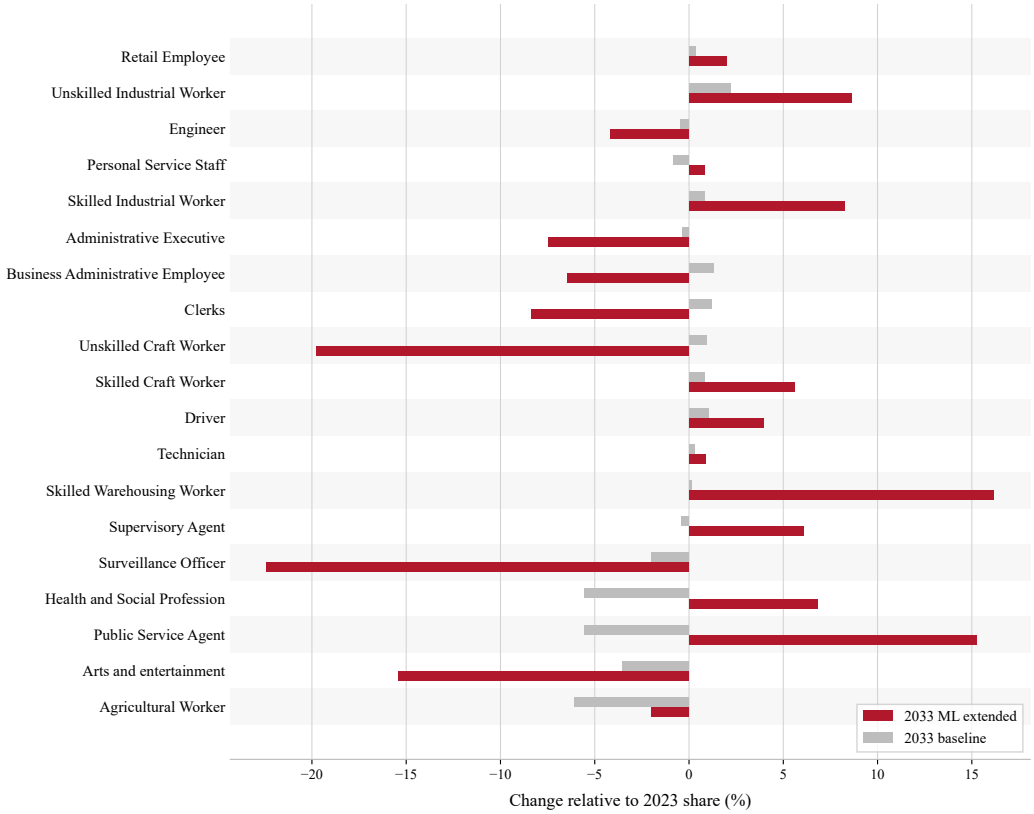
Figure E.2 isolates the contribution of leader information by subtracting the baseline forecast from the leader forecast. This difference exceeds 10% of the initial share in absolute value for five occupations: it is positive for public service agents, and skilled warehousing workers, and negative for surveillance officers, unskilled craft workers, and arts and entertainment. The additional information in leader workforce structures therefore changes both the magnitude and the occupational pattern of the projection, beyond what current and lagged industry shares alone predict.

Table E.2: Projected Occupational Structure in 2033

	Employment share (%)			Change 2023–2033 (pp)	
Occupation	2023	2033 baseline	2033 leaders	Total	Leader contribution
Unskilled industrial workers	9.04	9.25	9.82	+0.78	+0.58
Skilled industrial workers	7.51	7.57	8.13	+0.62	+0.56
Skilled warehousing workers	3.67	3.67	4.26	+0.59	+0.59
Skilled craft workers	6.22	6.27	6.57	+0.35	+0.30
Drivers	5.51	5.56	5.73	+0.22	+0.16
Retail employees	10.79	10.83	11.01	+0.22	+0.18
Public service agents	0.97	0.92	1.12	+0.15	+0.20
Supervisory agents	2.23	2.22	2.36	+0.14	+0.14
Personal service staff	8.47	8.40	8.54	+0.07	+0.14
Health and social professions	0.98	0.92	1.05	+0.07	+0.12
Technicians	4.53	4.54	4.57	+0.04	+0.03
Agricultural workers	0.58	0.54	0.57	−0.01	+0.02
Arts and entertainment	0.96	0.92	0.81	−0.15	−0.11
Engineers	8.48	8.44	8.13	−0.35	−0.32
Surveillance officers	1.92	1.88	1.49	−0.43	−0.39
Business administrative employees	6.81	6.90	6.37	−0.44	−0.53
Administrative executives	7.01	6.99	6.49	−0.52	−0.50
Clerks	6.45	6.53	5.91	−0.54	−0.61
Unskilled craft workers	6.36	6.42	5.10	−1.26	−1.32

Notes: Aggregate employment shares by two-digit PCS occupation for the firms and sectors covered by the sample. "2033 baseline" and "2033 leaders" are the projections of the baseline model and of the extended ML leader model, both estimated on 2010–2012–2022 and applied to 2021–2023, aggregated with 2023 industry employment weights and renormalized to sum to one; no flooring of negative cell predictions is applied. "Leader contribution" is the difference between the two projections. Occupations with less than 0.5% of employment in 2023 are omitted.

Figure E.1: Projected Occupational Share Changes, 2023–2033, Relative to Initial Shares



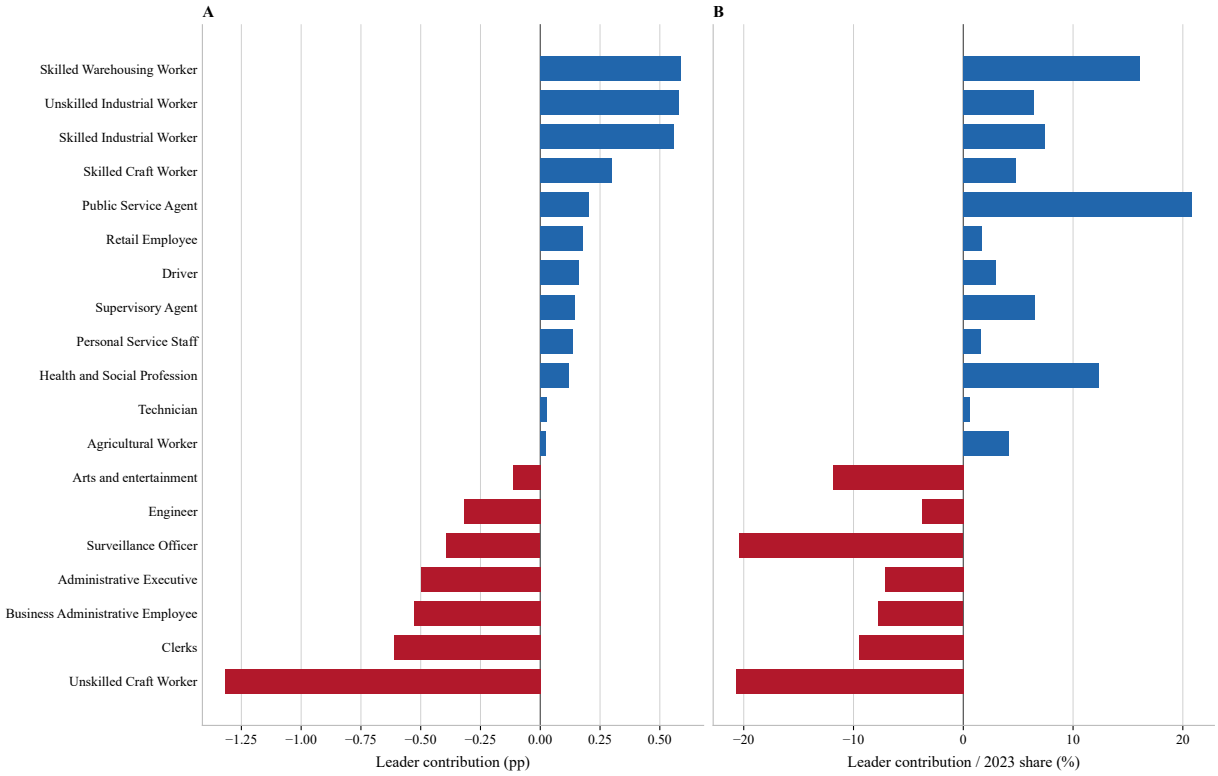
Notes: Same forecasts as Figure 8, expressed as $100(s_{2033} - s_{2023})/s_{2023}$ for each occupation, where s_{2033} is the predicted 2033 share of the extended machine-learning model (red) or of the baseline model (light gray) and s_{2023} the realized 2023 share. Occupations are in the same order as in Figure 8. The eight occupations with 2023 shares below 0.5% are excluded to limit the sensitivity of relative changes to small initial shares.

The sectoral composition of industrial occupations. The projected gains for industrial workers should not be read as a forecast of reindustrialization. Industry employment weights are held fixed, and much of the employment in these occupations is recorded in service firms. Table E.3 shows that manufacturing's share of skilled industrial workers falls from 50% in 2015 to 47% in 2023 and 43% in the projection. For unskilled industrial workers, the corresponding shares are 21%, 16%, and 13%. The gains in their aggregate occupational shares thus accompany a continued shift toward employment in services.

This distinction reflects the range of activities covered by the occupational categories. Skilled industrial workers include workers in public works, mechanical assembly, and industrial maintenance; unskilled industrial workers include handling, sorting, packing, and shipping workers.[33] The results are consistent with a growing role for service firms in supplying industrial

[33]Qualified warehouse keepers and forklift operators belong to the separate category of skilled warehousing workers (PCS 65), whose employment share also increases in the projection.

Figure E.2: Contribution of Leader Information to the 2033 Projection



Notes: Difference between the extended machine-learning forecast and the baseline forecast of the 2033 share of each occupation. Panel A reports the difference in percentage points, Panel B the difference divided by the realized 2023 share of the occupation, in percent. Occupations are ordered by the contribution in Panel A in both panels. Blue bars indicate a higher projected share under the leader model than under the baseline; red bars indicate a lower projected share. The eight occupations whose 2023 share is below 0.5 percent of employment are excluded from both panels. The same forecasts underlie Figure 8.

and logistics activities, potentially through outsourcing. The economic pattern is an increase in the employment share of industrial occupations within services, rather than an expansion of manufacturing's share of employment.

Technology exposure and the leader contribution. The exposure comparisons help characterize the occupational shifts anticipated by leaders. Table E.4 relates the incremental leader contribution to six technology measures. For extended ML leaders, the contribution over 2015–2023 is negatively associated with GPT exposure and with suitability for machine learning ($t = -2.34$ and -2.36), while the other four associations are not significant at the 5% level. In the 2033 projection, the contribution is negatively associated with AI and GPT exposure ($t = -2.39$ and -2.36), while its positive associations with computerization and automation risk are weaker and not significant ($t = 1.56$ and 1.81). Webb's AI exposure is unrelated to the contribution at both horizons ($|t| \leq 0.4$). Leader information therefore contributes to the contrast documented in the

Table E.3: Where Industrial Workers Are Employed

	Skilled industrial workers			Unskilled industrial workers		
Employing sector	2015	2023	2033	2015	2023	2033
Manufacturing	50.4	46.6	42.7	21.4	16.2	12.9
Construction	12.1	11.1	10.7	1.9	1.8	1.6
Trade and transport	4.7	5.8	5.0	11.6	11.0	8.7
Other services	31.2	34.3	39.4	65.0	70.9	76.7
Agriculture, mining and utilities	1.6	2.2	2.2	0.1	0.1	0.1
Share of total employment (%)	7.1	7.5	8.1	8.6	9.0	9.8

Notes: Distribution of each occupation's employment across employing sectors, in percent; columns sum to 100. 2015 and 2023 are observed; 2033 is the projection of the extended ML leader model with 2023 industry weights. Manufacturing is NAF 15–37, construction NAF 45, trade and transport NAF 50–52 and 60–64, agriculture, mining and utilities NAF 01–14 and 40–41, and other services comprise hotels and restaurants, finance, real estate, and business services (NAF 55 and 65–74); the sample excludes education, health, and personal services. The last row gives the occupation's share of total employment in the sample.

main text: projected gains tend to favor occupations exposed to older forms of automation, while losses tend to fall on occupations more exposed to AI.

This pattern is specific to the way leaders are identified. For simple productivity leaders, the projected contribution is negatively associated with computerization and automation risk ($t = -4.52$ and -2.85), and its association with AI exposure is not statistically significant at 5%. Selecting firms for their ability to anticipate occupational structures therefore yields a different projected pattern from selecting them on productivity alone.

Table E.4: Leader Contribution to Predicted Changes and Automation Indices

		2015–2023		2023–2033	
Leader model	Index	R^2	t-stat	R^2	t-stat
ML	Computerization	0.190	−2.37	0.054	1.17
ML	Automation risk	0.016	−0.62	0.080	1.44
ML	AI exposure (Felten)	0.078	−1.43	0.168	−2.20
ML	GPT exposure	0.154	−2.09	0.175	−2.26
ML	AI exposure (Webb)	0.030	0.86	0.004	0.30
ML	Suitability for ML	0.397	−3.97	0.048	−1.10
ML extended	Computerization	0.012	−0.54	0.092	1.56
ML extended	Automation risk	0.001	0.14	0.120	1.81
ML extended	AI exposure (Felten)	0.121	−1.82	0.192	−2.39
ML extended	GPT exposure	0.185	−2.34	0.188	−2.36
ML extended	AI exposure (Webb)	0.001	−0.13	0.005	0.35
ML extended	Suitability for ML	0.189	−2.36	0.049	−1.11
Productivity	Computerization	0.038	0.97	0.460	−4.52
Productivity	Automation risk	0.169	2.21	0.253	−2.85
Productivity	AI exposure (Felten)	0.342	−3.53	0.074	1.39
Productivity	GPT exposure	0.308	−3.27	0.028	0.83
Productivity	AI exposure (Webb)	0.085	−1.49	0.143	2.00
Productivity	Suitability for ML	0.201	−2.46	0.112	−1.74
Productivity extended	Computerization	0.039	0.99	0.120	−1.81
Productivity extended	Automation risk	0.172	2.23	0.002	0.20
Productivity extended	AI exposure (Felten)	0.347	−3.57	0.047	−1.09
Productivity extended	GPT exposure	0.313	−3.31	0.109	−1.72
Productivity extended	AI exposure (Webb)	0.086	−1.51	0.201	2.46
Productivity extended	Suitability for ML	0.202	−2.46	0.380	−3.83

Notes: Each row reports a regression of the leader contribution on the indicated index. In the 2015–2023 columns the dependent variable is the leader-minus-baseline difference in the predicted change in the occupation employment share from 2015 to 2023, in percentage points, from the model estimated on 2002–2004–2012; in the 2023–2033 columns it is the leader-minus-baseline difference in predicted 2033 shares (equivalently, in the predicted change from 2023 to 2033), in percentage points, from the model estimated on 2010–2012–2022. Computerization is the probability from Frey and Osborne (2017); Automation risk is the index from Nedelkoska and Quintini (2018); AI exposure is the AIOE index of Felten et al. (2021); GPT exposure is the share of tasks exposed to large language models from Eloundou et al. (2024) (GPT-4 rating, β definition); AI exposure (Webb) is the AI exposure score of Webb (2020); Suitability for ML is the index of Brynjolfsson et al. (2018). The three US-based AI indices are mapped from SOC to ISCO-08 and to two-digit PCS with the same crosswalks as the AIOE index. Regressions are weighted by occupation employment (2015 weights for the 2015–2023 columns, 2023 weights for the 2023–2033 columns). Occupation 44 (Clergy) is excluded, leaving 26 occupations.

F Forecast Accuracy by "Fast" and "Slow" Industry Groups

This appendix repeats the out-of-sample assessment of Table 5 within groups of industries. We rank the 49 industries by five measures of early technology adoption and split them into a fast and a slow group in four different ways, as described in the text below and in the notes of Table F.1. Table F.1 and Figure F.1 compare the forecast of the 2023 occupation structure of each group with the realized one (aggregated across industries in this group), and Table F.2 reports the error over the industry-by-occupation cells of each group. For the group occupation structure the extended ML model has a lower error than the baseline in all 40 groups, with ratios between 0.46 and 0.93, and a lower error than a naive forecast that assumes constant occupational shares in 37 of them; errors are larger in the fast groups for almost every ranking and split. At the industry by occupation level the extended ML model beats the baseline in 25 of the 40 groups.

Measures of technology adoption. The first three measures use US online job postings data provided by Lightcast (2026) in 2010–2014. We focus on six-digit NAICS industry. For each of the 1,057 industries, we compute the share of postings that ask for newly emerging non-AI skills or for newly emerging AI skills following the definition in Jaumotte et al. (2026) in 2010–2014. In addition, we also apply the definition of AI skills from the Stanford AI Index Report 2026 (Lightcast and Stanford Institute for Human-Centered Artificial Intelligence, 2026). We map these shares to our 49 French industries by chaining the 2017 NAICS codes to 2012 NAICS with the Census concordance and from there to the NACE Rev. 1.1 divisions. The share of a French industry is the employment weighted average of the shares of the US industries that map into it, with the 2012 County Business Patterns employment of each US industry times its crosswalk weight as the weight. US industries without County Business Patterns employment, 3 percent of postings, receive a weight proportional to their postings. About 72 percent of the US employment behind the postings lands in the 49 industries, the remainder being health, education, public administration, agriculture and personal services that our French data do not cover. The last two measures use the automation patents of Mann and Püttmann (2023) by sector of use, which are reported by four-digit SIC industry and which we map to the 49 industries with employment weighted SIC to NAICS to NACE crosswalks. 91 percent of the automation patents granted in 2000-2014 land in the 49 industries. The level measure is the number of automation patents per year in 2010-2014 and the growth measure is that number relative to the number per year in 2000-2009. The 2000-2009 and 2010-2014 levels rank the industries almost identically, with a rank correlation of 0.996 and the same groups under every split, so one row stands for both. The three skill measures agree with each other more than with the patent measures, and the growth measure is close to unrelated to the level measure.

Splits. Each ranking orders the 49 industries from fastest to slowest adoption. The median split puts the 24 highest ranked industries in the fast group and the other 25 in the slow group. The

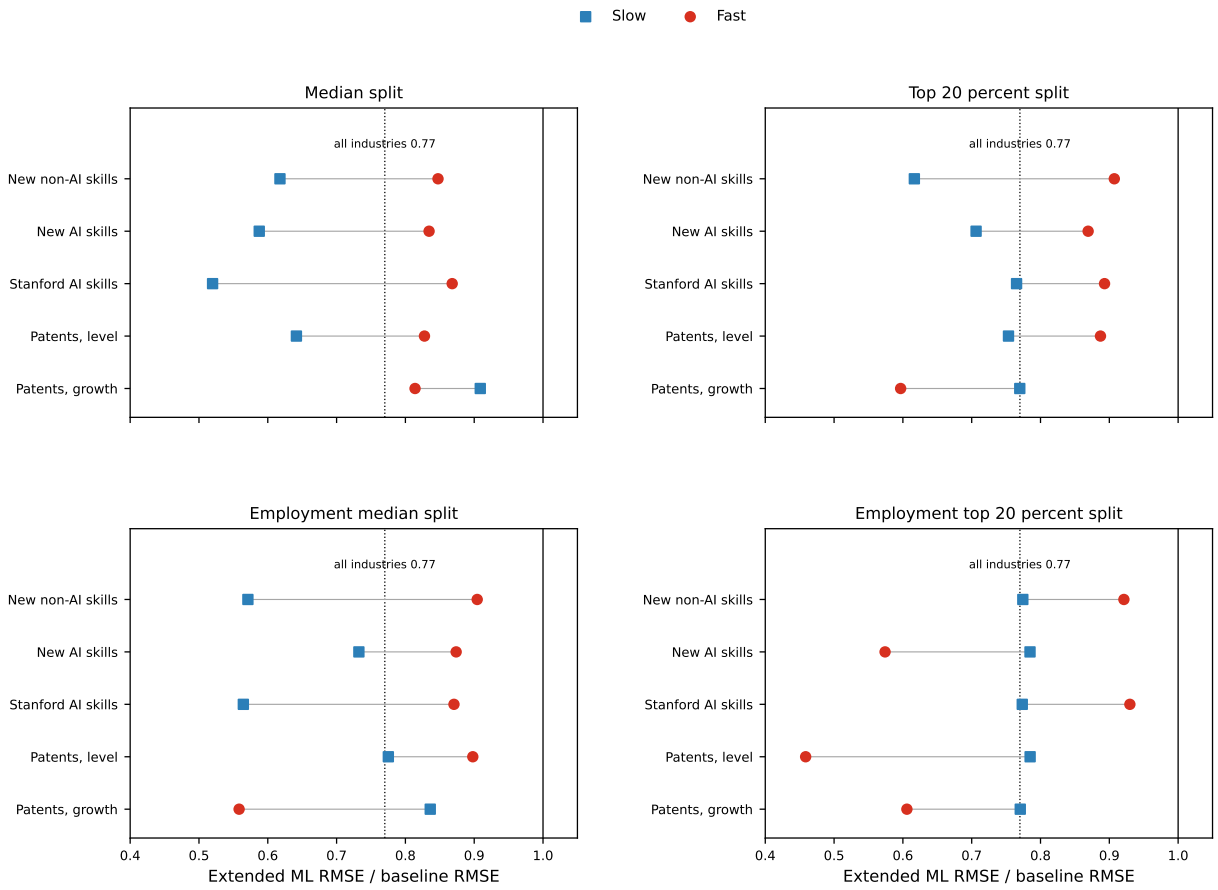
top 20 percent split puts the 10 highest ranked in the fast group. Because industries differ greatly in size, these two splits can produce fast groups with very different employment, from 7 to 75 percent of 2015 employment. The employment median split therefore adds industries to the fast group in rank order until it holds about 50 percent of 2015 employment, and the employment top 20 percent split until it holds about 20 percent. The main text reports the employment median split, for which the fast groups hold between 36 and 51 percent of employment and between 8 and 19 industries.

Table F.1: Out-of-Sample Forecast Error of the Group Occupation Structure, by Industry Group

	Fast group						Slow group					
Ranking	Ind.	Emp.	ML ext	Base	Naive	Ratio	Ind.	Emp.	ML ext	Base	Naive	Ratio
Panel A. Median split												
New non-AI skills	24	73	6.24	7.37	8.13	0.85	25	27	4.86	7.86	7.14	0.62
New AI skills	24	75	6.22	7.46	8.23	0.83	25	25	4.28	7.29	6.26	0.59
Stanford AI skills	24	59	7.51	8.65	9.34	0.87	25	41	3.41	6.56	5.71	0.52
Patents, level	24	71	6.13	7.40	7.69	0.83	25	29	4.90	7.64	7.46	0.64
Patents, growth	24	73	6.03	7.41	7.99	0.81	25	27	5.54	6.09	6.36	0.91
Panel B. Top 20 percent split												
New non-AI skills	10	43	8.60	9.48	10.32	0.91	39	57	4.02	6.53	6.42	0.62
New AI skills	10	58	6.98	8.03	8.55	0.87	39	42	4.63	6.55	6.82	0.71
Stanford AI skills	10	7	12.12	13.57	11.82	0.89	39	93	4.97	6.50	7.15	0.77
Patents, level	10	56	6.74	7.60	8.09	0.89	39	44	5.23	6.94	7.21	0.75
Patents, growth	10	19	6.51	10.90	8.85	0.60	39	81	5.45	7.08	7.51	0.77
Panel C. Employment median split												
New non-AI skills	16	47	8.30	9.17	9.98	0.90	33	53	3.70	6.47	6.19	0.57
New AI skills	8	51	7.45	8.52	9.04	0.87	41	49	4.36	5.96	6.29	0.73
Stanford AI skills	19	49	8.06	9.26	9.67	0.87	30	51	3.46	6.13	6.10	0.56
Patents, level	9	50	7.24	8.06	8.58	0.90	40	50	4.97	6.41	6.67	0.78
Patents, growth	19	36	3.89	6.97	6.31	0.56	30	64	6.44	7.70	8.22	0.84
Panel D. Employment top 20 percent split												
New non-AI skills	6	5	16.63	18.05	15.81	0.92	43	95	5.05	6.53	7.18	0.77
New AI skills	7	16	8.15	14.20	10.54	0.57	42	84	5.39	6.86	7.48	0.78
Stanford AI skills	16	13	12.81	13.77	12.76	0.93	33	87	5.05	6.52	7.19	0.77
Patents, level	7	14	6.74	14.69	10.29	0.46	42	86	5.58	7.10	7.68	0.78
Patents, growth	13	20	6.15	10.15	8.54	0.61	36	80	5.48	7.11	7.51	0.77
All 49 industries	49	100	5.08	6.59	7.16	0.77						

Notes: Each row splits the 49 industries into a fast and a slow group by the indicated ranking and split and compares the forecast of the 2023 occupation structure of each group with the realized one. Predicted within-industry shares are aggregated over the group's industries with 2015 employment weights, realized shares with 2023 weights, and the error is the root mean squared difference over the 26 occupations. Ind. is the number of industries in the group and Emp. its share of 2015 employment in percent. ML ext, Base and Naive are the out-of-sample root mean squared errors of the extended ML model, the baseline model without leaders and the no-change forecast that carries the 2015 shares forward, in share units times 1,000; Ratio is ML ext over Base. Rankings order the 49 industries from fastest to slowest technology adoption: the share of US job postings in 2010–2014 that ask for newly emerging non-AI skills, for newly emerging AI skills and for the AI skills of the Stanford AI Index, mapped from six-digit NAICS to the French industries with employment weights, and the automation patents of Mann and Püttmann (2023) by sector of use, as patents per year in 2010–2014 (Patents, level) and as the 2010–2014 rate over the 2000–2009 rate (Patents, growth); the 2000–2009 and 2010–2014 levels give the same groups under every split, so one row stands for both. Splits: the median split puts the 24 highest ranked industries in the fast group, the top 20 percent split the 10 highest ranked; the employment median and employment top 20 percent splits add industries in rank order until the fast group holds about 50 and about 20 percent of 2015 employment. Occupation 44 (Clergy) is excluded.

Figure F.1: Extended ML error relative to the baseline error, by industry group



Notes: Ratio of the out-of-sample root mean squared error of the extended ML model to that of the baseline model for the group occupation structure, within each group of Table F.1. Red dots are the fast groups, blue squares the slow groups of the same ranking and split. The dotted line marks the ratio on all 49 industries together and the solid line a ratio of one.

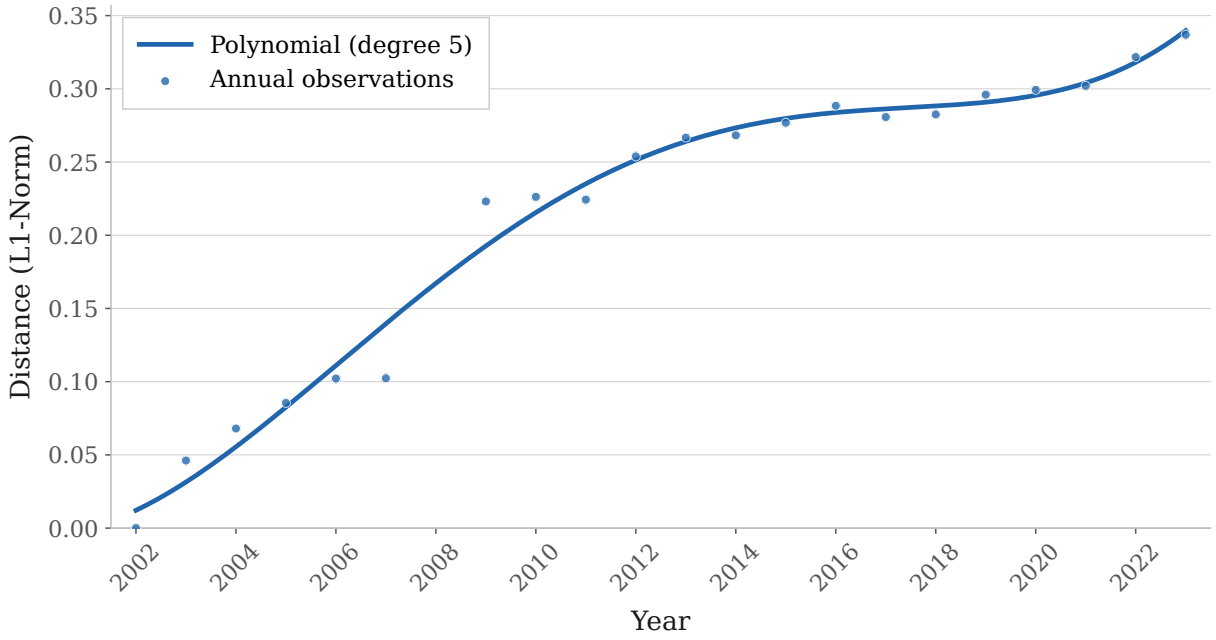
Table F.2: Out-of-Sample Forecast Error by Industry and Occupation Cell, by Industry Group

	Fast group						Slow group					
Ranking	Ind.	Emp.	ML ext	Base	Naive	Ratio	Ind.	Emp.	ML ext	Base	Naive	Ratio
Panel A. Median split												
New non-AI skills	24	73	12.67	13.01	12.54	0.97	25	27	11.58	13.92	12.25	0.83
New AI skills	24	75	12.70	13.03	12.55	0.98	25	25	11.34	13.98	12.18	0.81
Stanford AI skills	24	59	12.03	11.72	12.24	1.03	25	41	12.89	15.26	12.78	0.84
Patents, level	24	71	12.29	12.89	12.12	0.95	25	29	12.61	14.13	13.26	0.89
Patents, growth	24	73	12.67	14.33	13.08	0.88	25	27	11.56	9.80	10.60	1.18
Panel B. Top 20 percent split												
New non-AI skills	10	43	11.59	11.56	12.24	1.00	39	57	12.94	14.41	12.62	0.90
New AI skills	10	58	10.53	12.60	11.71	0.84	39	42	14.52	14.13	13.42	1.03
Stanford AI skills	10	7	16.58	15.66	14.65	1.06	39	93	12.01	13.07	12.28	0.92
Patents, level	10	56	9.47	12.01	11.15	0.79	39	44	15.28	14.70	13.94	1.04
Patents, growth	10	19	15.93	18.73	15.43	0.85	39	81	11.41	11.65	11.67	0.98
Panel C. Employment median split												
New non-AI skills	16	47	13.77	12.54	13.08	1.10	33	53	10.98	13.89	11.87	0.79
New AI skills	8	51	10.31	12.88	11.67	0.80	41	49	14.25	13.66	13.25	1.04
Stanford AI skills	19	49	11.61	11.59	11.96	1.00	30	51	13.08	14.70	12.93	0.89
Patents, level	9	50	9.24	12.37	11.24	0.75	40	50	14.90	14.11	13.58	1.06
Patents, growth	19	36	14.18	17.11	14.33	0.83	30	64	11.25	10.51	11.28	1.07
Panel D. Employment top 20 percent split												
New non-AI skills	6	5	15.06	13.66	12.23	1.10	43	95	12.22	13.24	12.48	0.92
New AI skills	7	16	11.82	17.51	13.15	0.67	42	84	12.49	12.28	12.33	1.02
Stanford AI skills	16	13	15.73	15.01	14.45	1.05	33	87	11.81	12.99	12.15	0.91
Patents, level	7	14	8.60	16.89	12.11	0.51	42	86	12.90	12.57	12.52	1.03
Patents, growth	13	20	16.09	18.59	15.58	0.87	36	80	11.26	11.56	11.55	0.97
All 49 industries	49	100	12.38	13.27	12.46	0.93						

Notes: Each row splits the 49 industries into a fast and a slow group by the indicated ranking and split and compares the forecast of the 2023 occupation share of every industry by occupation cell in the group with the realized one. The error is the root mean squared difference over the cells, with each industry weighted by its 2015 employment share within the group. Ind. is the number of industries in the group and Emp. its share of 2015 employment in percent. ML ext, Base and Naive are the out-of-sample root mean squared errors of the extended ML model, the baseline model without leaders and the no-change forecast that carries the 2015 shares forward, in share units times 1,000; Ratio is ML ext over Base. Rankings order the 49 industries from fastest to slowest technology adoption: the share of US job postings in 2010–2014 that ask for newly emerging non-AI skills, for newly emerging AI skills and for the AI skills of the Stanford AI Index, mapped from six-digit NAICS to the French industries with employment weights, and the automation patents of Mann and Püttmann (2023) by sector of use, as patents per year in 2010–2014 (Patents, level) and as the 2010–2014 rate over the 2000–2009 rate (Patents, growth); the 2000–2009 and 2010–2014 levels give the same groups under every split, so one row stands for both. Splits: the median split puts the 24 highest ranked industries in the fast group, the top 20 percent split the 10 highest ranked; the employment median and employment top 20 percent splits add industries in rank order until the fast group holds about 50 and about 20 percent of 2015 employment. Occupation 44 (Clergy) is excluded.

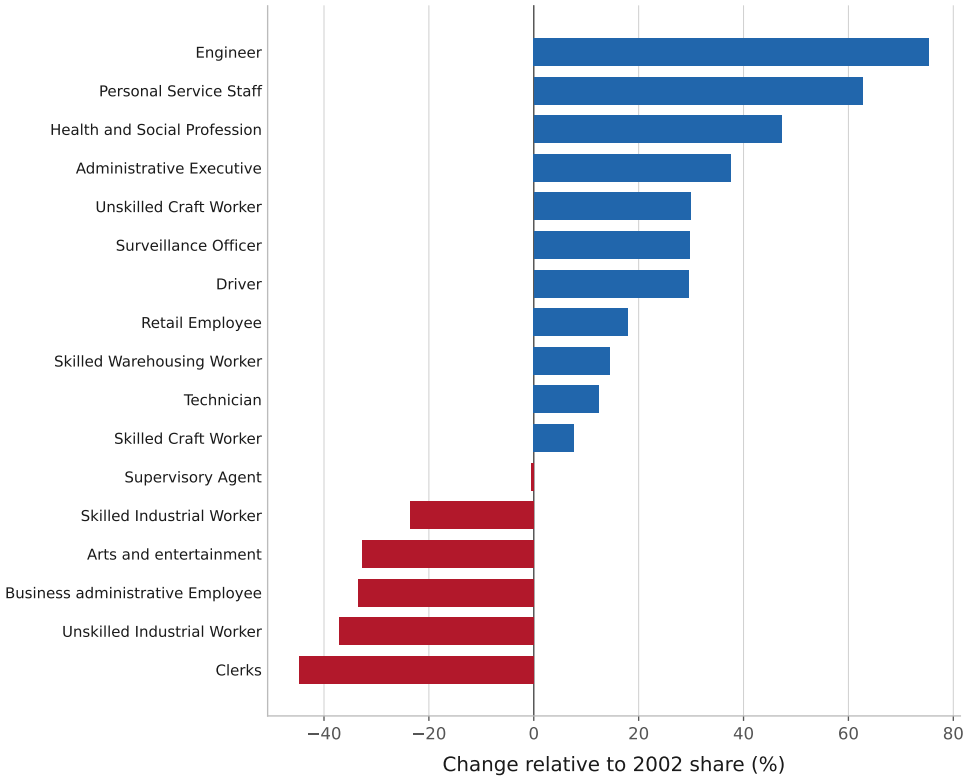
G Additional Figures and Tables

Figure G.1: Change in the Occupational Structures over Time measured by L_1 Norm



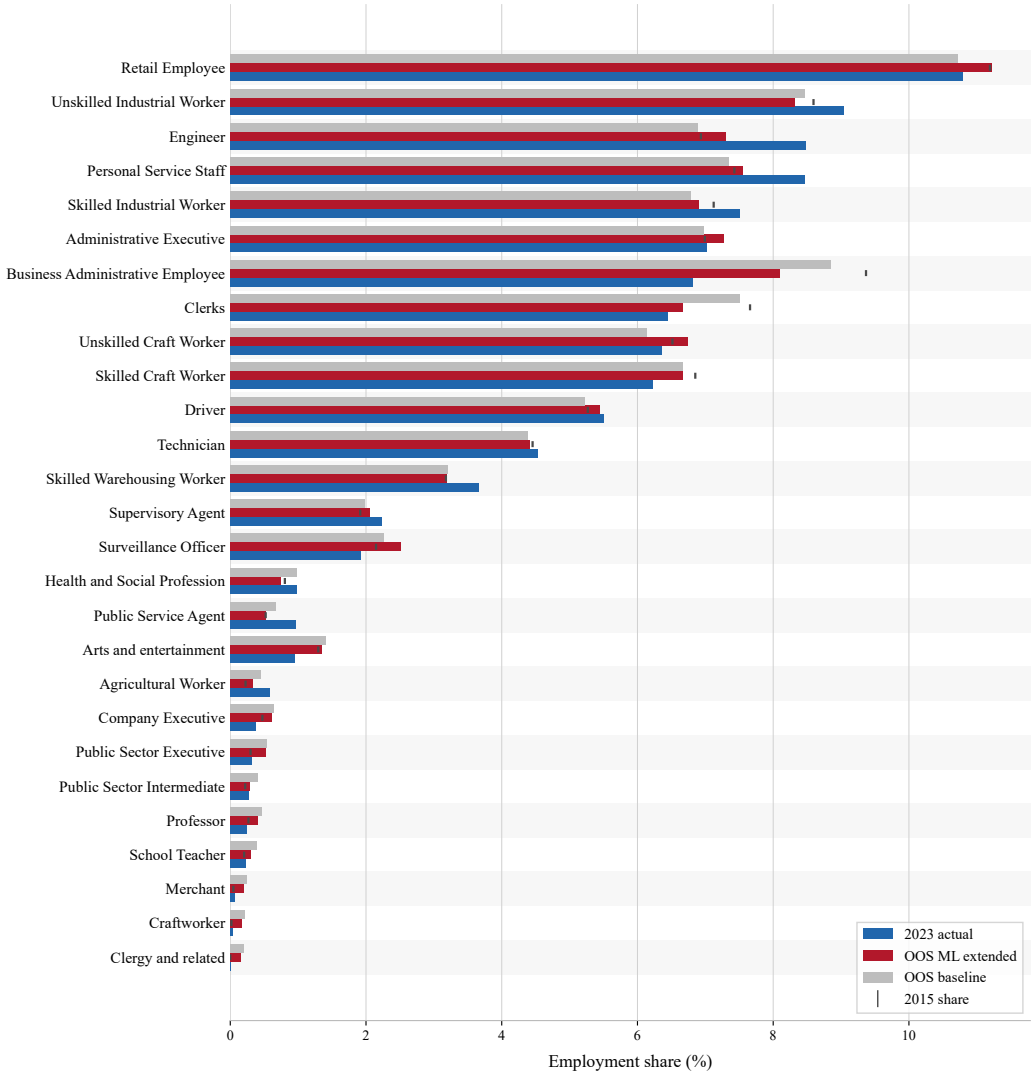
Notes: The figure shows the distance between occupational structures the respective year relative to the year 2002, measured by the L_1 norm, for two-digit PCS occupations. The solid blue line is a fifth-degree polynomial fitted by ordinary least squares to the annual L_1 norm distances from the 2002 base year over 2002–2023. The time variable was rescaled as $t = (\text{year} - 2002)/21$ to improve numerical stability. The polynomial provides a smooth approximation of the observed trend. The year 2008 was excluded due to missing data.

Figure G.2: Percent Changes in the 2-digit Occupational Structures between 2002 – 2023



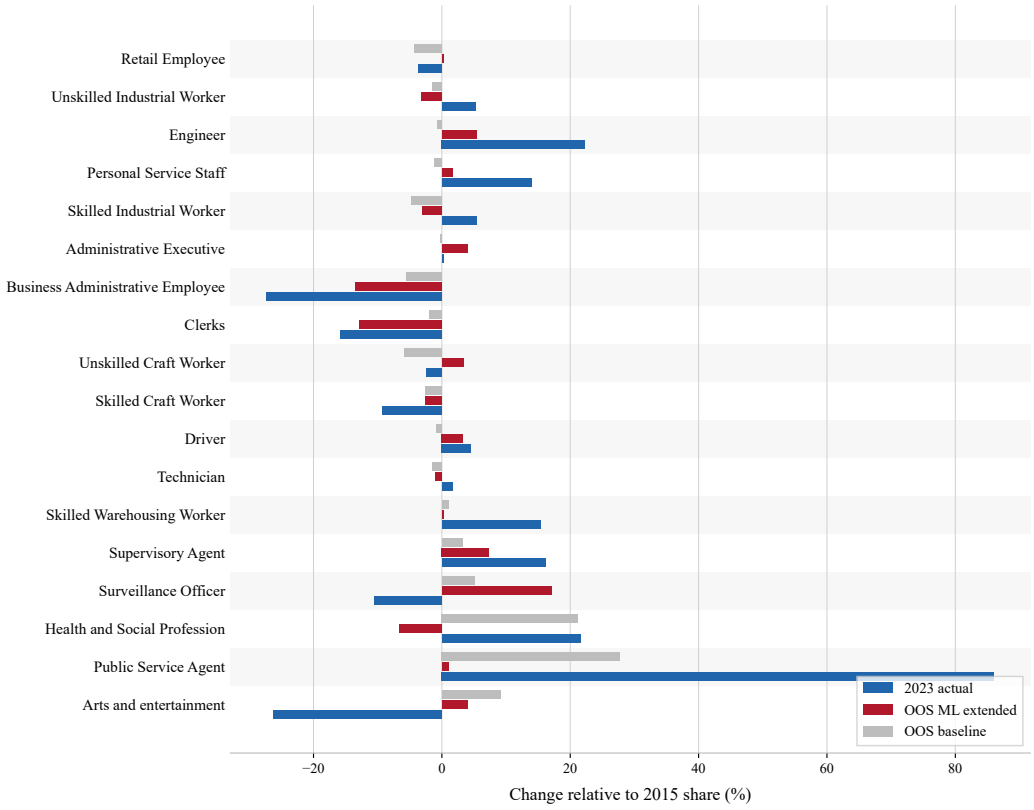
Notes: Same changes as Figure 2, expressed as $(S^o_{2023} - S^o_{2002})/S^o_{2002}$ for each occupation o. Recall that S^o_t is the national employment share of occupation o in year t. Occupations are sorted by the size of the change. The ten occupations whose 2002 share is below 0.5 percent of total employment are excluded because the ratio is not informative for them. Their changes in percent of the 2002 share are: agricultural workers $+711$, public sector executives $+545$, public sector intermediate professions $+263$, craftworkers $+101$, public service agents $+101$, merchants $+84$, school teachers $+1$, professors -2, company executives -18, clergy and related -88.

Figure G.3: Out-of-sample forecast of the 2023 occupational structure by occupation



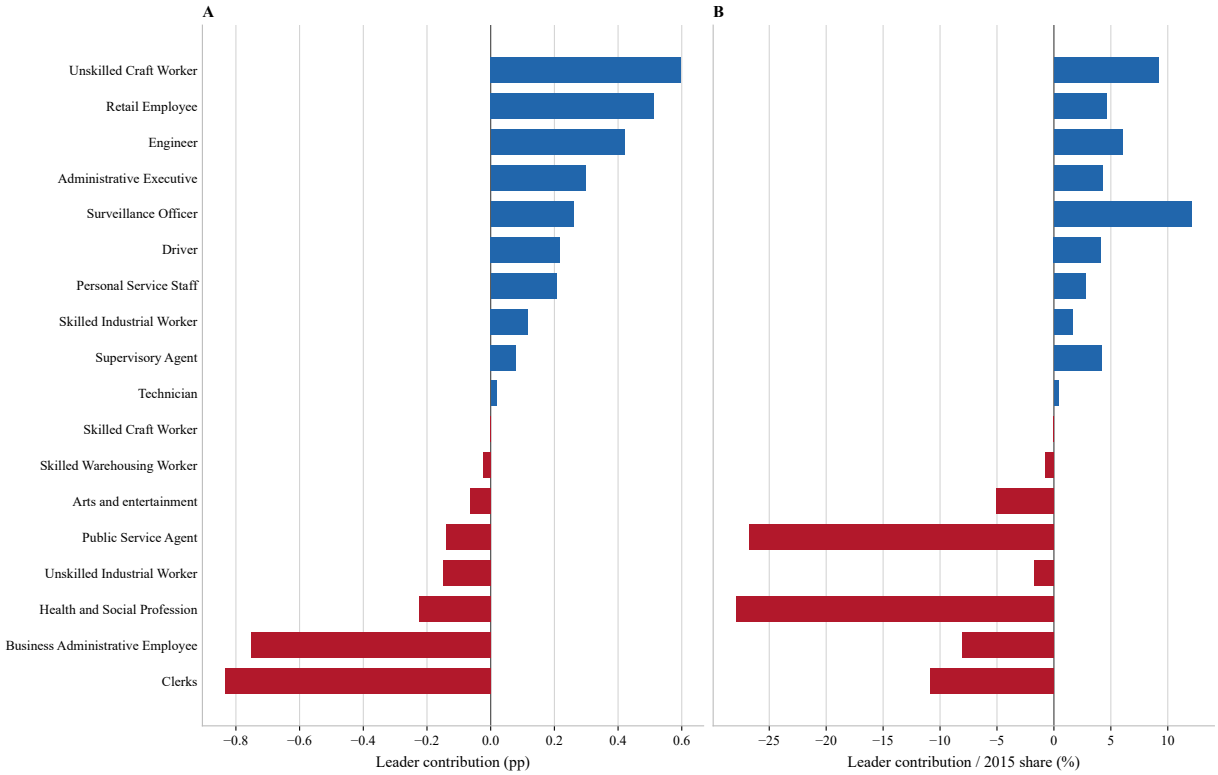
Notes: Realized 2023 occupation structure (blue) and the out-of-sample forecasts of the extended machine-learning model (red) and of the baseline model (light gray) that underlie Panel B of Table 5. The thin gray bar on top of the blue bar marks the 2015 share of the occupation. Forecasting coefficients are estimated on 2002–2004–2012, leaders are identified in 2015 and covariates are taken from 2013 and 2015. Shares aggregate industry–occupation cells to national occupation shares using 2015 industry employment shares as weights. Occupations are ordered by their 2023 share.

Figure G.4: Change from 2015 to 2023 in percent of the 2015 occupation share: realized and out-of-sample forecasts



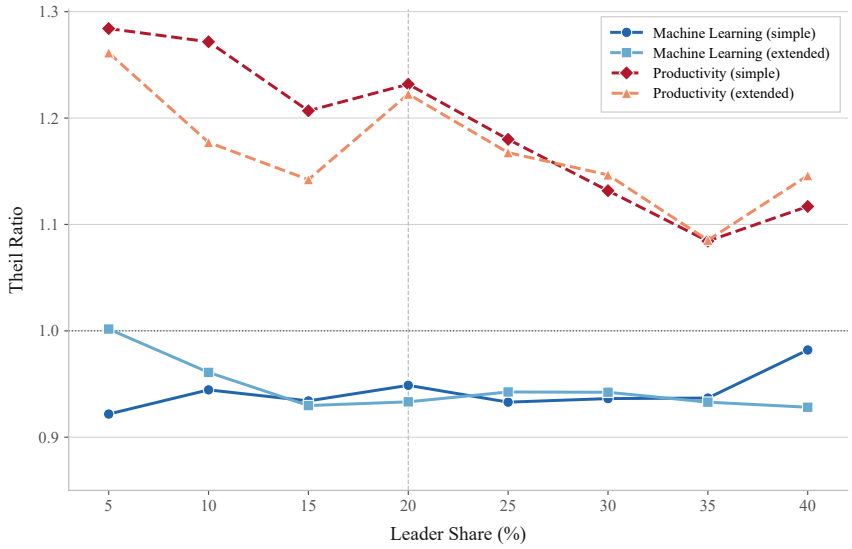
Notes: Same shares as Figure G.3, expressed as $(s_{2023} - s_{2015})/s_{2015}$ for each occupation, where s_{2023} is the realized 2023 share (blue), the extended machine-learning forecast (red) or the baseline forecast (light gray) and s_{2015} the realized 2015 share. Occupations are in the same order as in Figure G.3. The nine occupations whose 2015 share is below 0.5 percent of employment are excluded.

Figure G.5: Leader contribution to the out-of-sample 2023 forecast



Notes: Difference between the extended machine-learning forecast and the baseline forecast of the 2023 share of each occupation, out of sample. Panel A reports the difference in percentage points, Panel B the difference divided by the realized 2015 share of the occupation, in percent. Occupations are sorted by Panel A in both panels. The nine occupations whose 2015 share is below 0.5 percent of employment are excluded from both panels. The same forecasts underlie Figure G.3.

Figure G.6: Out-of-sample Assessment with changing Leader Shares: industry-occupation forecast



Notes: Out of sample Theil ratios as a function of the cumulative employment share used to define leaders within each two-digit industry. The baseline specification uses a 20% leader share. Values below one indicate improved forecasting accuracy relative to the baseline model without leaders.

Table G.1: Occupation-mix Distances and Overlap of Leader Sets

A. Within-industry occupational distance			
		All firms	
	Retrained leaders	in 2015	in 2023
Retrained 2015 leaders		0.164	0.145
Baseline rule 2015 leaders	0.144	0.149	0.157
Surviving 2004 leaders	0.257	0.169	0.188
B. Turnover of 2004 ML leaders			
Share surviving to 2015 sample	68.6%		
Of survivors, still retrained 2015 leaders	31.0%		

Notes: Panel A shows the distance of the occupational employment distributions by industry of the set of firms in the respetive row-column combination. The distance is measured through the $L2$ norm: Each entry is the square root of the 2015-employment-weighted mean, across two-digit industries, of the squared difference in occupation shares. *Retrained 2015 ML leaders* are the 2015 leading firms from a random forest trained on 2013–2015–2023 data. The *baseline rule 2015 ML leaders* are leading firms in 2015 from the 2002–2004–2012 random forest applied to 2015 firm characteristics. *Surviving 2004 ML leaders* are ML leaders in 2004 which are still present in the 2015 sample. The columns *All firms in 2015* and *2023* are based on the industry-level occupation mix in those years. Panel B focuses on *Surviving 2004 ML leaders* and their share among all 2004 ML leaders. It also shows the (unweighted) overlap of surviving 2004 ML leaders with 2015 retrained ML leaders.