Montobbio, Fabio; Staccioli, Jacopo; Virgillito, Maria Enrica; Vivarelli, Marco

**Working Paper**

# Labour-saving automation and occupational exposure: A text-similarity measure

LEM Working Paper Series, No. 2021/43

**Provided in Cooperation with:**
Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies

*Suggested Citation:* Montobbio, Fabio; Staccioli, Jacopo; Virgillito, Maria Enrica; Vivarelli, Marco (2021) : Labour-saving automation and occupational exposure: A text-similarity measure, LEM Working Paper Series, No. 2021/43, Scuola Superiore Sant'Anna, Laboratory of Economics and Management (LEM), Pisa

This Version is available at:
https://hdl.handle.net/10419/259538

INSTITUTE
OF ECONOMICS

Scuola Superiore
Sant'Anna

LEM | Laboratory of Economics and Management

Institute of Economics
Scuola Superiore Sant'Anna

Piazza Martiri della Libertà, 33 - 56127 Pisa, Italy
ph. +39 050 88.33.43
institute.economics@sssup.it

# LEM
# WORKING PAPER SERIES

## Labour-saving automation and occupational exposure: a text-similarity measure

Fabio Montobbio *†‡
Jacopo Staccioli *§
Maria Enrica Virgillito §*
Marco Vivarelli *¶||

* Department of Economic Policy, Università Cattolica del Sacro Cuore, Milan, Italy
† BRICK, Collegio Carlo Alberto, Turin, Italy
‡ ICRIOS, Bocconi University, Milan, Italy
§ Institute of Economics, Scuola Superiore Sant'Anna, Pisa, Italy
¶ UNU-MERIT, Maastricht, The Netherlands
|| IZA, Bonn, Germany

**2021/43** **November 2021**
**ISSN(ONLINE) 2284-0400**

# Labour-saving automation and occupational exposure: a text-similarity measure

Fabio Montobbio[*,†,‡] Jacopo Staccioli[*,§]

Maria Enrica Virgillito[§,*] Marco Vivarelli[*,¶,||,**]

22nd November 2021

**Abstract**

This paper represents one of the first attempts at building a *direct* measure of occupational exposure to robotic labour-saving technologies. After identifying robotic and labour-saving robotic patents retrieved by Montobbio et al. (2022), the underlying 4-digit CPC definitions are employed in order to detect functions and operations performed by technological artefacts which are more directed to substitute the labour input. This measure allows to obtain fine-grained information on tasks and occupations according to their similarity ranking. Occupational exposure by wage and employment dynamics in the United States is then studied, complemented by investigating industry and geographical penetration rates.

**JEL classification:** O33, J24.

**Keywords:** Labour-Saving Technology, Natural Language Processes, Labour Markets, Technological Unemployment.

[*]Department of Economic Policy, Università Cattolica del Sacro Cuore, Via Necchi 5 – 20123 Milano, Italy.

[†]BRICK, Collegio Carlo Alberto, Piazza Arbarello 8, 10122 Torino, Italy.

[‡]ICRIOS, Bocconi University, Via Röntgen 1, 20136 Milano, Italy.

[§]Institute of Economics, Scuola Superiore Sant'Anna, Piazza Martiri della Libertà 33, 56127 Pisa, Italy.

[¶]Maastricht Economic and Social Research Institute on Innovation and Technology, UNU-MERIT, Boschstraat 24, 6211 AX Maastricht, The Netherlands.

[||]Forschungsinstitut zur Zukunft der Arbeit GmbH (IZA), Schaumburg-Lippe-Strasse 5-9, 53113 Bonn, Germany.

[**]To whom correspondence should be addressed: ✉ marco.vivarelli@unicatt.it

# 1 Introduction

*Robots are coming!* Such a statement has been the mantra in the last recent years, with the perception that *"This time is really different"* (Brynjolfsson and McAfee, 2012, 2014; Ford, 2015). A large literature on the effects of the new wave of automation on human labour blossomed since then. Indeed, the pervasiveness of such new technological artefacts has been one of the most relevant aspects pushing for troublesome scenarios; among the most radical, Frey and Osborne (2017) suggests that 47% of total US employment is associated with occupations that are potentially automatable, a very much debated figure and downward revised by further estimations (Arntz et al., 2016). The recent empirical evidence tends however to agree that low- and medium-skilled workers, mainly executing routinised tasks, are particularly at risk (Acemoglu and Restrepo, 2018, 2019, 2020a; Autor and Dorn, 2013; Frey and Osborne, 2017). At the same time, while some papers find a negative impact on employment and wages, systematic evidence on the labour market impact of robotic technologies remains elusive (Calvino and Virgillito, 2018; Mondolo, 2021).

The unfolding of robotic impact on labour markets, in terms of occupations and wages, has been mainly estimated in the literature by two alternate methods. The first method is based on experts judgement on a subset of occupations, expanded over the entire occupational structure by a classifier-system algorithm (e.g. Frey and Osborne, 2017; Nedelkoska and Quintini, 2018). The second approach has been leveraging on robotic adoption at the sectoral level, relying on the International Federation of Robotics dataset and looking at the impact on local labour markets (e.g. Acemoglu and Restrepo, 2018, 2019, 2020a).

Currently, a direct measure of human substitutability and occupational exposure, ideally based on the effective functions and operations executed by labour saving (LS) technologies, is still absent (see Section 2). We contribute to this literature and provide a direct link between human tasks and machine functions and, as a result, quantify occupational exposures to LS innovation in robotics. In doing so, we build a new measure of similarity between the textual description of the tasks performed by an occupation and the functions performed by observed robotic LS innovations.

First, we leverage on the identification of robotic LS technologies by means of natural language processing on robotic patents (Montobbio et al., 2022) and we perform a task-based textual match between the descriptions of technological classifications (so called CPC codes) attributed to robotic LS patents and the O*NET dictionary of occupations. The match exploits a cosine-similarity matrix that measures the proximity of the two dictionaries of words. The first result of our study

is therefore the construction of a direct measure of similarity between a dictionary of technological LS functions and a dictionary of human-based functions. This is a methodological advancement to measure proximity between humans and machines, and allows to recover a direct measure of exposure.

In a second step, we aggregate tasks into occupations and recover a measure of exposure of each task and related occupations to robotic LS technologies. We find that the distribution of the similarity scores across tasks and occupations is very skewed, with high similarity events being quite rare, given the underlying heterogeneity between the two text corpora. Nonetheless, restricting the analysis to the top-decile of the similarity distribution, around 8.6% of the overall US employed workforce (approximately 12.6 million jobs) is at risk of substitutability. The most affected occupations are "Material Moving Workers", "Vehicle and Mobile Equipment Mechanics, Installers, and Repairers", "Other Production Occupations". Logistics and standardised service activities are those most exposed to LS technologies, in line with the evidence that among top-owners of LS patents, Amazon and UPS stand out (Montobbio et al., 2022).

To validate our methodology, we perform a robustness analysis by replicating the text similarity exercise between robotic LS patents' full-texts and the same O*NET task descriptions. The patent-task match, when LS functions are not taken into consideration (through CPC codes descriptions, see above), correctly pinpoints those occupations developing new innovative robotic technologies and their systems of adoption (e.g. Robotics Engineers; Robotics Technicians). This result reinforces the goodness of our procedure because it shows that it tells apart substitutability detected via more prevalent functions in LS patents from complementarity detected via the match with the entire patent text.

Then, we link the similarity measure to the actual US labour market in terms of occupations and wages. We match our data to the Occupational Employment and Wage Statistics (OEWS) from US Bureau of Labor Statistics for 8-digit SOC occupations (1999-2019). LOWESS estimates present a monotonically negative relationship between occupational exposure and both (i) wage level and growth, and (ii) employment level and growth. Remarkably, the expected U-shaped pattern (Acemoglu and Autor, 2011) is not recovered, neither in wages, nor in occupational growth. In other words, cutting-edge robotic innovative efforts look to be directed towards the weakest and cheapest segment of the labour market and not versus the middle one. Finally, the geographical breakdown across US states shows that the Rust Belt area, the region surrounding the Great Lakes experiencing industrial decline, and the South-East area, with higher prevalence of African-American communities, record the largest employment shares of occupations that are particularly exposed to robotic LS technologies.

The remainder of the paper is organised as follows: Section 2 discusses the literature and evidence available; Section 3 presents the datasets used and Section 4 the adopted methodology; Section 5 shows and discusses our results, presenting task and occupational exposure, labour market, industry and geographical penetration rates. Section 6 concludes the paper.

## 2 Literature review

As briefly mentioned in the introduction, the unfolding of robotic applications on labour markets, in terms of occupations and wages, has been mainly estimated in the literature by two alternate methods. The first method is the one initiated by Frey and Osborne (2017), which constructed an automation probability starting from experts judgement on a subset of 70 occupations, and then expanding the evaluation over the entire occupational structure by means of a classifier-system algorithm. Experts were asked about the probability of automating some particular human functions. This approach has been then employed in Nedelkoska and Quintini (2018) to study 32 OECD countries using the PIAAC dataset and downward revised by Arntz et al. (2016).

The second approach has been leveraging on robotic adoption at the sectoral level, relying on the International Federation of Robotics dataset and looking at the impact on local labour markets. This is the route taken by Acemoglu and Restrepo (2018, 2019, 2020a) which generally predict that a higher number of robots per employee decreases wages and occupations for low-wage workers. However, cross-country studies at the industry level do find a positive impact of robotics adoption on labour productivity, and less clear-cut evidence on employment reduction. For instance, while Chiacchio et al. (2018) find results very consistent with Acemoglu and Restrepo (2020a), both Graetz and Michaels (2018) and Dauth et al. (2017) conclude that robots do not significantly reduce total employment, although they do reduce the low-skilled workers' employment share, particularly in manufacturing.

Shifting to studies using firm-level data, results are conflicting. Domini et al. (2020), using robotic adoption or, alternatively, imported capital equipment, does not detect labour expulsion, but rather employment growth. Interestingly enough, in some studies the positive employment impact at the firm level appears entirely due to the so-called "business stealing effect" – i.e. innovative adopters gain market shares at the expense of non-innovators (Dosi and Mohnen, 2019) – since negative employment impacts emerge once non-adopters and sectoral aggregates are taken into account (see Acemoglu et al., 2020b; Koch et al., 2021).

Other studies, using longitudinal data and a more comprehensive measure of

embodied technological change (which includes robots) do find a labour-saving effect of new technologies (see Barbieri et al., 2018; Pellegrino et al., 2019). Such contradictory results, out of different levels of aggregations, are not new to scholars in the economics of innovation (Clark et al., 1981; Freeman and Soete, 1987) which identify alternate effects of technical change on employment via a series of compensation mechanisms balancing the initial direct LS effects of mechanisation and automation (Dosi et al., 2021; Piva and Vivarelli, 2018; Simonetti et al., 2000; Van Roy et al., 2018; Vivarelli, 1995, 2015). Indeed, the level of analysis, whether sectoral or establishment/firm, produces different signs on the underlying relationship (Calvino and Virgillito, 2018).

More recent papers have focused on artificial intelligence, the purportedly newcomer disruptive technology, often blamed to have a strong LS impact on white-collar jobs, more related to service activities. Felten et al. (2021), which refines the measure proposed in Felten et al. (2018), links the Electronic Frontier Foundation dataset (EFF), within the AI Progress Measurement initiative, with O*NET (abilities). A direct matching between 10 AI selected scopes of application (abstract strategy games, real-time video games, image recognition, visual question answering, image generation, reading comprehension, language modelling, translation, and speech recognition) and human abilities is conducted. The matching is performed by crowdsourcing a questionnaire to gig workers at Amazon's Mechanical Turk (mTurk) web service. To 2,000 mTurkers residing in the United States the questions administered ask whether, for each of the 52 abilities listed in the O*NET, they believe that the AI application is related to or could be used for. The study reports higher AI exposure for white-collar workers. However, the measure is silent about any direct replacement or complementarity effect.

Webb (2020) proposes a direct measure of exposure via co-occurrence of verb-noun pairs in the title of AI patents and O*NET tasks. However, titles of patents are hardly informative of the underlying functions executed by the technological artefact and restricting to verb-noun pairs has high likelihood of false positives. The measure of exposure is not constructed in terms of overall similarity of the two corpora but rather in terms of the relative frequency of occurrence of the elicited pair in AI titles versus the remaining titles of non-AI patents. Moreover, the proposed methodology does not allow to distinguish labour-saving from labour-augmenting technologies.

Acemoglu et al. (2020a) look at AI exposed establishments and their job posts using Burning Glass Technologies data, which provide wide coverage of firm-level online job postings, linked to SOC occupational codes. In order to account for the degree of firm-level AI exposure, three alternative measures are employed, namely the ones put forth by Brynjolfsson et al. (2018), Felten et al. (2021) and Webb (2020).

Not surprisingly, considering the still relatively niche adoption, no clear effects at the industry and occupational level are detected, while recomposition toward AI-intensive jobs is spotted. In addition, they do not find evidence of any direct complementarity between AI job posts and non-AI jobs, hinting therefore at a prevalent substitution effect and workforce recomposition, rather than a productivity-enhancing effect of AI adoption.

The closest analysis to our own is the one performed in Kogan et al. (2021) which constructs a text-similarity measure between the corpus of so called *breakthrough innovations*, according to the methodology devised in Kelly et al. (2018), and the fourth edition of the Dictionary of Occupation Titles (DOT). The measure is constructed to allow for time variability by keeping constant the textual content similarity but summing it for each defined breakthrough innovation at each time step, exploiting patent information over the period 1850-2010. Breakthrough innovations, identified as the distance between backward and forward similarity of each filed patent compared to the existing stock of patents, are by no means ex-ante defined as being in nature labour-saving ones. In addition, the way the measure is built reflects more the dynamics of breakthrough innovations according to their emergence along subsequent technological revolutions, quite akin to the findings in Staccioli and Virgillito (2021), rather than the actual penetration of these technologies in the labour market. Therefore, what the measure captures is more the clustering of technologies under mechanisation in the first period of analysis, followed by automation and the ICT phase. They find that most exposed occupations lost in terms of wages and employment level, and that over time white-collar workers became relatively more exposed compared to blue-collar ones. However, it is not clear whether the results are reflecting more long-run dynamics in technological and structural change rather than actual similarity between patents and occupations. Indeed, the within patent-occupation text-similarity is kept constant over time.

The Kogan et al. (2021)'s measure has been applied in Autor et al. (2020) interested in devising the entry of new work titles along the historical records of the so called Census Alphabetical Index of Occupations (CAI), an index listing all new work-title entries. The authors define complementary-technologies those patents matched with the CAI text (new job titles), and labour-saving technologies the ones linked to the DOT text (existing job titles). The paper documents the increasing entry of white-collar middle-paid occupations in the period 1940-1980, while since 1980 new jobs have been concentrating in both high-educated and low-educated services. Another application of the Kogan et al. (2021)'s measure has been done with reference to I4.0 patents in Meindl et al. (2021), matching in this case the patent text corpus with the "detailed work activities" (DWAs) section of the O*NET.

According to their results, financial and occupational professions are more exposed to I4.0 patents compared to non I4.0 patents.

## 3 Data description

The first dataset used in this study is represented by the O*NET, a primary source of occupational information, largely employed in the literature studying the actual content of the workplace activities (Handel, 2016). It provides details for the US occupational structure at the 8-digit level. The O*NET content model allows to gather information on a series of attributes of the world of work, namely executed tasks, task ratings (in terms of importance, relevance, and frequency), abilities, education, training and experience, knowledge, skills, work activities, work context, work styles.

Among the many available descriptors, the detailed Task Statements descriptor contains specific definitions of the tasks performed by each 8-digit occupation, while the Task Statement Ratings allows to gather information on the actual importance, relevance and frequency of each task, being each occupation composed by a multiplicity of tasks, also variable across occupations. The definitions of *Core* or *Supplemental* tasks synthesise the numerical rating scores, as detailed below.

Take as an example the occupation `19-3011.00` defining *"Economists"*. The latter perform 11 core tasks, some of them more specific to the occupation in itself (e.g. task `7537`: *"Develop economic guidelines and standards and prepare points of view used in forecasting trends and formulating economic policy"*); some other tasks are less occupational specific (e.g. task `7542`: *"Supervise research projects and students' study projects"*); some others considered to be supplemental tasks (e.g. task `20051`: *"Provide litigation support, such as writing reports for expert testimony or testifying as an expert witness"*). Such type of granular information will be the basis to construct the dictionary of words defining human functions.

The second dataset employed is the Occupational Employment and Wage Statistics (OEWS) retrieved from the US Bureau of Labor Statistics. Such dataset allows to analyse the evolution of the employment dynamics, excluding self-employed, and it is directly linkable to the O*NET dataset via the SOC occupational codes at full-digit. In addition the dataset allows to recover information on the average and median nominal wages for each 8-digit SOC category.

Fig. 1 is a snapshot of the US occupational structure in 2019, showing occupational categories aggregated at 2-digit levels (22 codes, excluding military specific occupations) (top panel) and the evolution over the last two decades in terms of employment shares (bottom panel). In 2019, the largest occupational share (14%) is populated by *"Office and Administrative Support"* workers; *"Healthcare Practition-*

*ers"* and *"Technical, Business and Financial Operations, Management"* both stand at 6%, while *"Life, Physical, and Social Science"* are less than 1%. The snapshot tells about the remarkable de-industrialisation of the US economy, with prevalence of administrative operations but also of service activities related to the satisfaction of social needs, such as *"Food Preparation and Serving Related"*, *"Transportation and Material Moving"*, while *"Production"* is relegated to the fifth position with less than 7% of the US workforce. The bottom panel presents the change in employment shares: the most growing occupation is *"Healthcare Support"* with almost a 100% increase as employment share, followed by *"Business and Financial Operators"*, *"Computer and Mathematical"* occupations. In general, a negative relationship between employment share growth in the last twenty years and employment share levels in 2019 is detectable, with those occupations recording the highest shares also experiencing strong contraction, like *"Office and Administrative Support"* or contraction, like *"Sales and Related"*. *"Production"* workers record the largest decline in employment share (-4%), a further confirmation of the accelerated de-industrialisation process in the US.

Fig. 2 presents the corresponding information for nominal median wages by occupational categories at the 2-digit level. While an evident hierarchical distribution of median wages emerges (top panel), with managerial median wages five times higher than food preparation activities, the bottom panel presents the demeaned dynamics of median nominal wage growth. Albeit all occupations have experienced a generalised nominal wage growth, with a minimum 50% increase, top-growing occupations in terms of remuneration have been both top-paid and bottom-paid ones according to the 2019 wage level: take the opposite dynamics of managerial vs. legal activities, in top remuneration tiers, that in the first case experienced an almost doubling remuneration when compared to 1999 levels, and a 20% increase more than the average, while in the second case recorded the lowest median wage increase with respect to average one. Regarding the lower tiers, *"Food Preparation and Serving Related"* activities, the least-paid occupations in 2019, have likely experienced a wage growth over twenty years and a 10% wage increase higher than the average, but still maintain their relative position in the wage hierarchy. Such crystallised hierarchical structure informs about a more rigid than expected US labour market, wherein notwithstanding occupational changes in relative employment shares, the wage distribution across occupational categories is quite sticky over twenty years.

The third employed dataset is represented by the universe of USPTO patent applications in robotic technologies in the period 2009-2018, the last recent phase recording a steep increase in patenting activity in such field (cf. Fig. 3). As we shall

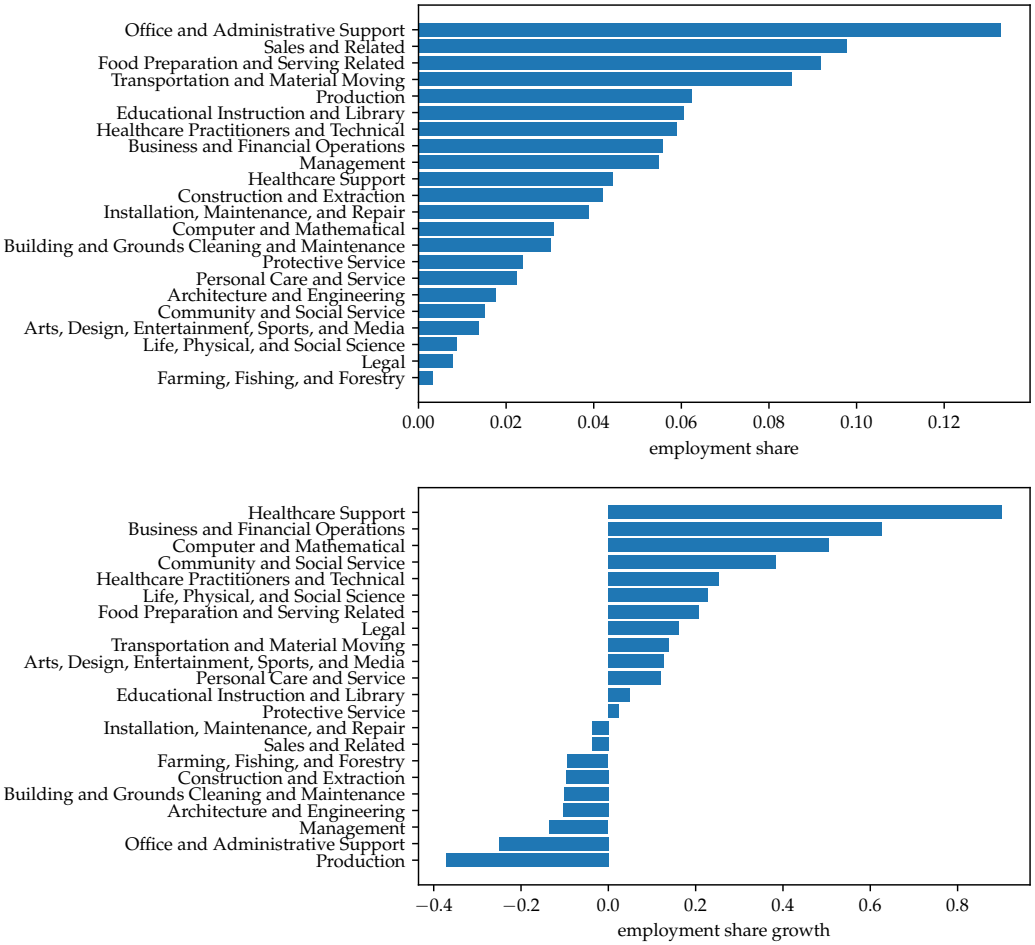


Figure 1: Employment shares in 2019 and share changes (1999-2019) for 2-digit SOC occupations.

describe below, both patent text corpus and the technological fields are taken into account (CPC classification at 4-digit level).

## 4 Methodology

In the present section, we develop the necessary methodology to build a new measure of similarity between the textual description of the tasks performed by an occupation and the functions performed by LS innovations. In particular, we leverage on the text similarity between the definitions of CPC (Cooperative Patent Classification) codes and the descriptions of tasks contained in the O*NET dictionary of occupations. Before delving into the methodological details of the text similarity measure that we devise (Section 4.2), it is useful to first summarise the relevant

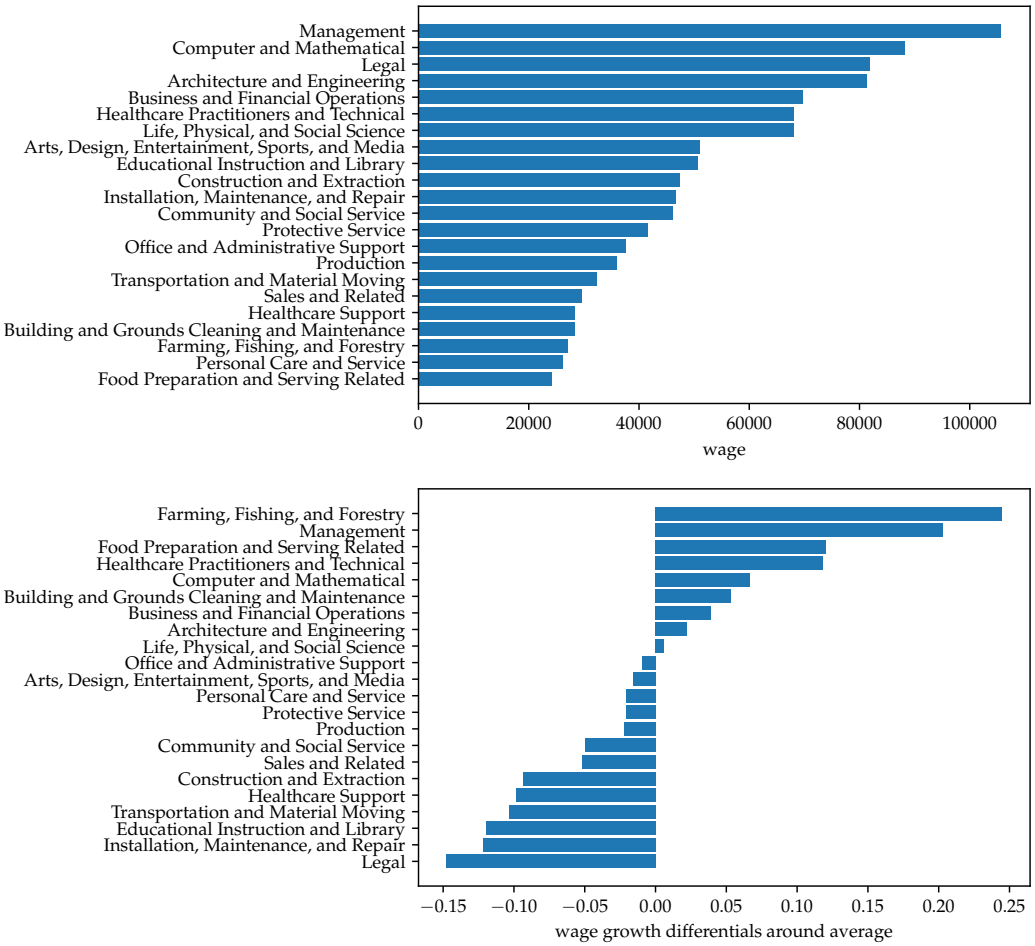


Figure 2: Nominal wage level in 2019 and demeaned nominal wage changes (1999-2019) for 2-digit SOC occupations.

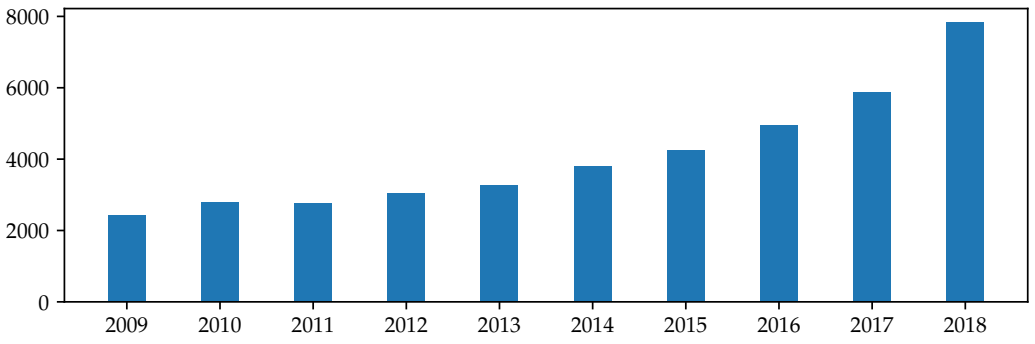


Figure 3: Number of USPTO robotic patents per year. Source: own elaboration based on Montobbio et al. (2022, Fig. 2).

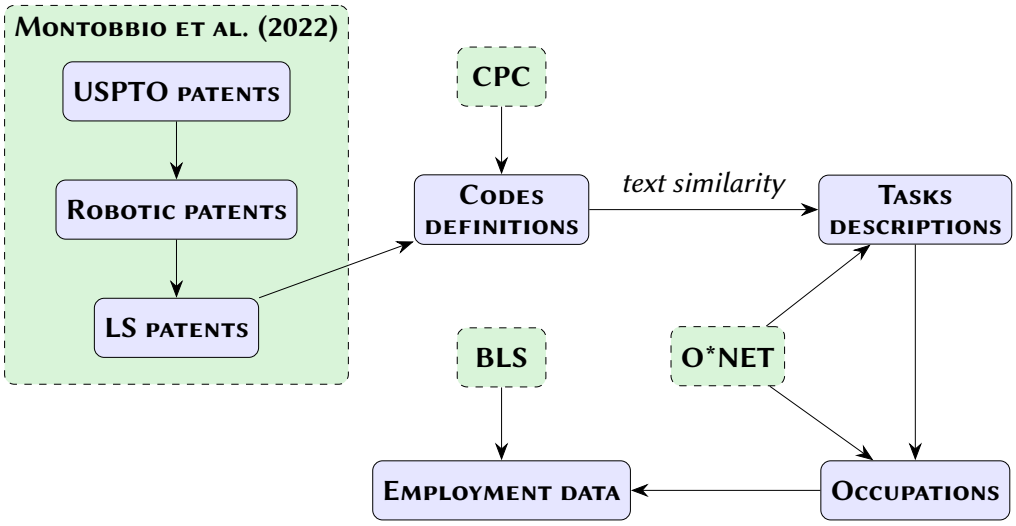


Figure 4: Flowchart of our methodology.

workflow of Montobbio et al. (2022) which brought about the set of LS patents which constitutes the starting point of the present analysis (Section 4.1). In Section 4.3 we explain how we map our measure of exposure from the task level to the occupation level. Our methodological workflow is synthesised by the flowchart in Fig. 4.

### 4.1 Discovery of labour-saving patents

The contribution of Montobbio et al. (2022) in the discovery of robotic LS patents unfolds along three methodological steps. First, patents which either directly or indirectly relate to robotics technology are singled out. Second, a procedure is implemented in order to detect the underlying LS heuristics and pinpoint the set of explicitly LS patents. Finally, most relevant CPCs in robotic LS patents vis-à-vis sheer robotic patents are identified. A brief technical summary of the relevant workflow is presented in Appendix A.

### 4.2 Measuring exposure with text similarity

Within the scope of the present analysis, the technological content of LS patents is proxied by the official definitions of the relevant CPC codes. Patent publications are each assigned one or more classification terms indicating the subject to which the invention relates. The CPC system (of which we use the version `2019.08`) consists of 250,000+ distinct codes, organised according to a multi-level hierarchical structure. For the purpose of striking a fair balance between density of information and granularity, we focus on 4-digit codes, of which 671 are present.

In order to match the technological content of LS patents to occupations, we rely on the O*NET database (of which we use version `25.1`) which contains a thorough description of 19,231 distinct tasks, further aggregated into 923 8-digit SOC2018 occupations according to a weighing scheme detailed in Section 4.3. In the following, we aim at measuring the pairwise text similarity of the 671 CPC codes' definitions and the 19,231 tasks' descriptions.

From the methodological point of view, we adopt the so-called *bag-of-words* model and we measure textual proximity between CPC definitions and task descriptions by means of *cosine similarity* (see e.g. Aggarwal, 2018). The bag-of-words model entails the representation of text as a *multiset* of underlying words, which disregards any grammar structure and the order in which terms appear, but keeps their multiplicity. The underlying assumption is that CPC-task pairs whose text consists of the very same words, possibly repeatedly, are more associated to one another than pairs which share few common words, or their frequency is negligible.

Each piece of text, either a CPC definition or a task description, first undergoes a preprocessing step in which words are *stemmed* to their morphological root and so-called stop-words, i.e. tokens that are overly common in English (such as `'a'`, `'the'`, `'if'`...) and do not convey any useful information to our analysis, are removed. Each text is then transformed into a vector of frequencies of the underlying words. The number of vector components reflects the common dictionary of terms across the two whole corpora. In other words, all vectors belong to the same vector space, whose dimension equals the number of distinct words in the common dictionary. The similarity of each CPC-task pair is then quantified as the cosine of the angle between the two underlying vectors.

As opposed to simply counting the occurrences of each word in each body of text, we adopt the customary *tf-idf* (term frequency–inverse document frequency) term-weighting scheme for computing the relevant frequencies, according to the following definition.

**Definition 1** *Let $D$ be a collection of documents $d$, each composed of an ensemble of terms $t$ from a dictionary $T$. The tf-idf measure of term $t$ appearing in document $d$ is defined as follows:*

$$\begin{aligned}
\text{tf-idf}(t,d,D) &:= \text{tf}(t,d)\cdot\text{idf}(t,D) && \forall\, d\in D,\ \forall\, t\in T,\\
\text{tf}(t,d) &:= \mathbf{1}_d(t) = \begin{cases} 1 & \text{if } t\in d \\ 0 & \text{otherwise} \end{cases} && \forall\, d\in D,\ \forall\, t\in T,\\
\text{idf}(t,D) &:= \log\left(\frac{|D|}{|\{d\in D : t\in d\}|}\right) && \forall\, t\in T.
\end{aligned}$$

The associated $|D| \times |T|$ document term matrix $\mathcal{D}^D$ is an array of *tf-idf* measures for all documents $d$ in the generic collection $D$ and for all terms $t$ in the relevant dictionary $T$. In other words,

$$\mathcal{D}^D_{d,t} = \text{tf-idf}(t, d, D), \quad \forall\, d \in D,\ \forall\, t \in T.$$

The tf-idf statistic reflects how important a specific term is to a certain document, compared to other documents in the collection. The tf-idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus which mention that word. This helps to adjust for the fact that some words appear more frequently in general.

Extending the reasoning to the corpus level, we construct two *document-term* matrices, $\mathcal{D}^{CPC}$ and $\mathcal{D}^{TASK}$, whose rows contain the aforementioned tf-idf frequency vectors, for each CPC code definition and each task description, respectively. Both matrices are based on the dictionary of terms from CPC definitions, namely the smaller between the two collections, which consists of 2309 terms. Therefore $\mathcal{D}^{CPC}$ has dimension $671 \times 2309$ and $\mathcal{D}^{TASK}$ has dimension $19231 \times 2309$.

Finally, we construct the cosine similarity matrix $\mathcal{S}$ containing the cosine similarity score between all pairs of row vectors from the document-term matrices $\mathcal{D}^{CPC}$ and $\mathcal{D}^{TASK}$ according to the following definition.

**Definition 2** *Given two vectors $X, Y \in \mathbb{R}^{|T|}$, their cosine similarity is defined as the cosine of the angle between them, which is also equal to the inner product of the same vectors normalised to unit length, as follows:*

$$\cos(X, Y) := \frac{X \cdot Y}{\|X\|\ \|Y\|} = \frac{\sum_{t=1}^{|T|} x_t y_t}{\sqrt{\sum_{t=1}^{|T|} x_t^2}\sqrt{\sum_{t=1}^{|T|} y_t^2}}, \tag{♣}$$

*where $x_t$ and $y_t$ denote the components of vectors $X$ and $Y$, respectively, and $\|\cdot\|$ denotes the Euclidean norm.*

Since row vectors of document-term matrices are non-negative valued, their cosine similarity is bounded by the unit interval, i.e. $\cos(X, Y) \in [0, 1]$. Moreover, when term frequency is measured by tf-idf, the normalisation denominator in eq. (♣) is redundant and $\cos(X, Y) \equiv X \cdot Y$. Therefore, given document-term matrices $\mathcal{D}^{CPC}$ and $\mathcal{D}^{TASK}$, and extending the cosine similarity computation to the matrix level,

$$\mathcal{S} = \cos\left(\mathcal{D}^{CPC},\ \mathcal{D}^{TASK}\right) \equiv \mathcal{D}^{CPC}\left(\mathcal{D}^{TASK}\right)'.$$

| OCCUPATION | 11-1011.00 | | ··· | ··· | 53-7121.00 | |
|---|---|---|---|---|---|---|
| TASK / CPC | 8823 | 8824 | ··· | ··· | 12809 | 12810 |
| A01B | cos(A01B,8823) | cos(A01B,8824) | ··· | ··· | cos(A01B,12809) | cos(A01B,12810) |
| A01D | cos(A01D,8823) | cos(A01D,8824) | ··· | ··· | cos(A01D,12809) | cos(A01D,12810) |
| ··· | ··· | ··· | ··· | ··· | ··· | ··· |
| H05H | cos(H05H,8823) | cos(H05H,8824) | ··· | ··· | cos(H05H,12809) | cos(H05H,12810) |
| H05K | cos(H05K,8823) | cos(H05K,8824) | ··· | ··· | cos(H05K,12809) | cos(H05K,12810) |

Table 1: Architecture of the cosine similarity matrix $\mathcal{S}$.

The cosine similarity matrix $\mathcal{S}$ has dimension $671 \times 19231$, one row for each CPC code and one column for each task, and each cell contains the similarity score of the underlying CPC-task pair. In total, there exist 12,904,001 such pairs. The architecture of matrix $\mathcal{S}$ is represented in Table 1, where the task-occupation mapping, defined in the O*NET database itself, is also highlighted.

In order to make the cosine similarity matrix reflect the technological structure of the LS patents found by Montobbio et al. (2022), we multiply each row of $\mathcal{S}$ by the frequency of the associated CPC code in the whole set of LS patents. When a patent is assigned multiple CPC codes with the same 4-digit representation, each is taken into account separately. In other words, we filter the cosine similarity matrix by the distribution of technological codes across LS patents. In this way, we weigh the contribution of each CPC code to the occupational exposure of a certain task proportionately with how widespread the code appears in LS patents (cf. Table 5). Finally, in order to rank O*NET tasks by similarity score with the ensemble of CPC codes of LS patents, we compute column sums of matrix $\mathcal{S}$ across all CPC codes (rows). The result is a task similarity vector $TS$ containing a unique measure of aggregate similarity to each task. Given a column vector $C$ of frequencies of the 671 CPC codes among LS patents, vector $TS$ is defined simply as

$$TS = \mathcal{S}'C$$

where the usual matrix multiplication is intended. A ranking of tasks by (aggregate) similarity score is later presented in Table 2, where values have been rescaled between 0 and 1.

### 4.3 From tasks to occupations

So far we have measured the textual proximity of LS patents to each O*NET task, mediated by CPC codes, the result of which is stored in the task similarity vector $TS$ (cf. Section 4.2). In order to draw conclusions on the effect of LS technologies

upon employment, we need to further aggregate the similarity measure at the occupation level.

The O*NET database defines occupations as a collection of underlying tasks, distinguishing between *core* and *supplemental* tasks. This classification takes into account three distinct measures, namely *importance*, *relevance*, and *frequency*. In order to aggregate a task similarity measure to the relevant occupation in a sensible way, it is crucial to understand how the core/supplemental distinction is devised in the first place. Importance spans a range between 1 and 5, while relevance and frequency are both represented as percentages. Core tasks are deemed critical to the occupation; the criteria a task to be classified as core require that relevance $\geq 67\%$ and importance $\geq 3.0$. Supplemental tasks are deemed less relevant and/or important to the occupation; two sets of tasks are included in this category, namely tasks rated $\geq 67\%$ on relevance but $< 3.0$ on importance, and tasks rated $< 67\%$ on relevance, regardless of importance.

Taking the O*NET definition of core and supplemental tasks into account, we impute task similarity to occupations with the following weights:

$$\text{core}: \quad \frac{2/3}{\#\text{ tasks in the occupation}}$$

$$\text{supplemental}: \quad \frac{1/3}{\#\text{ tasks in the occupation}}$$

It is worth noticing that occupations differ in the number of constituent tasks. This warrants the need of rescaling the similarity contribution of each task accordingly, hence the denominator in the aforementioned weighting scheme. A ranking of occupations by (aggregate) similarity score is later presented in Table 3, where values have been rescaled between 0 and 1.

## 5 Results

In the following, we shall display our results. This section is organised as follows: Section 5.1 presents the distribution of the text similarity measure across tasks and occupations. Section 5.2, exploiting the match with the OEWS dataset, allows to understand the degree of penetration of technological exposure in terms of structure of the labour market, looking at employment and wages, and industry composition. Finally, Section 5.3 presents the geographical distribution of LS threats.

### 5.1 Task and occupational exposure

Table 2 shows the top-fifteen tasks in terms of the similarity score. Notably, the tasks presenting a higher similarity score regard human functions related to the

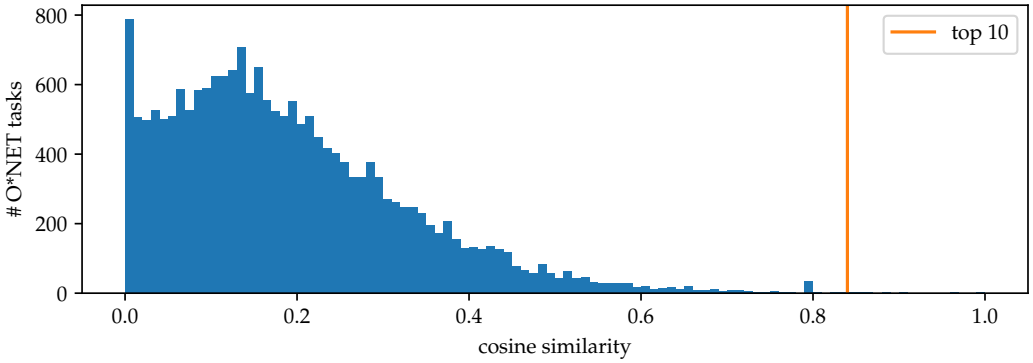


Figure 5: Distribution of cosine similarity with respect to tasks.

handling and moving of objects, materials, products, with the first top-three tasks being (i) *"Load materials and products into machines and equipment, or onto conveyors, using hand tools and moving devices"*; (ii) *"Move levers or controls that operate lifting devices, such as forklifts, lift beams with swivel-hooks, hoists, or elevating platforms, to load, unload, transport, or stack material"*; (iii) *"Position lifting devices under, over, or around loaded pallets, skids, or boxes and secure material or products for transport to designated areas"*. Such types of tasks are more prevalent in the logistics industry inasmuch they require activities like preparing boxes, packaging, sorting, and routing items.

Fig. 5 presents the distribution of the similarity measure across the entire set of 19,231 O*NET tasks, according to the Task Statements file. Given the wide spectrum of the covered tasks by the O*NET, it is not surprising to obtain an extremely skewed distribution, with events of high similarity being extremely rare. Indeed, cosine similarity values beyond 0.8 apply to a tiny fraction of tasks and at this stage seem to constitute more the exception rather than the norm. However, such an extreme value distribution is quite comforting in terms of reliability of the measure, the underlying information being quite sparse, inasmuch the probability of false positives is low and the overall accuracy of the measure high. Nonetheless, one might consider irrelevant or less informative the task domain, given that occupations are defined by not as a single, but rather as a set of distinct tasks. Most exposed tasks therefore might result to be largely not core in many occupations, in line with the evidence provided in Fig. 1, telling a picture of a labour market where administrative, office, and sales occupations represent the largest shares of employed workers.

However, such a handful of tasks, when aggregated into occupations, happens to be quite revealing of the direction of robotic LS efforts. Table 3 shows the corresponding top-twenty occupations most exposed to substitution exerted by some

form of automation or intelligent automation, obtained by aggregating tasks into occupations, as described in Section 4.3. Although the distribution by tasks is quite skewed, the very same information, once aggregated at the SOC level, allows the detection of a series of occupations at a strong exposure risk. Most exposed occupations are *"Industrial Truck and Tractor Operators"*; *"Maintenance Workers, Machinery"*; *"Machine Feeders and Offbearers"*; *"Packers and Packagers, Hand"*. Such occupations are clearly among the ones reporting the higher incidence of tasks with higher similarity scores. Browsing Table 3, it emerges a recurrent pattern of some specific macro-occupational groups, as evident by the presence of the 2-digit occupations *"Transportation and Material Moving"* (53), *"Installation, Maintenance, and Repair"* (49), *"Packaging and Filling Machine Operators and Tenders"* (51), all recurrently ranking in the top-twenty 8-digit most exposed occupations.

Fig. 6 presents the histogram of occupations by similarity. Albeit less skewed than the task histogram, it confirms that high similarity is a rare event, affecting only a relatively small fraction of the entire occupational structure. The top-twenty occupations with the highest similarity scores are indeed very few, and the mode of the distribution stands in the medium range of similarity, between 0.2 and 0.4. The area on the right of the orange bar identifies the range of the top-twenty most exposed occupations. But, what does it happen if we move the bar to the left?

Fig. 7 (left-panel) plots the quantile function of the similarity distribution in Fig. 6. Given the highly skewed pattern, up to the eighth decile of the similarity distribution there is a low range of variation, reaching the value of 0.4. Higher cosine similarity values, in the range $[0.6,\ 1]$ occur only after the inflection point located around the ninth decile. This non linear-relationship informs about the existence of a threshold level beyond which exposure dramatically increases while below such point, exposure is tamed. Such threshold behaviour has a twofold implication: from the one hand, high-exposure risk to substitutability affects a relatively tiny fraction of the entire occupational range, while from the other hand, whenever the risk is high, it swiftly accelerates, potentially leading to quite probable substitutability events.

Fig. 7 (right-panel), by using the O*NET-OEWS match, allows to recover the effective number of employed workers per each occupation at risk of substitutability. As expected, the number of replaceable employees dramatically drops when the similarity value increases: the top-decile of the similarity distribution, on the far-right, affects 8.6% of the employed working population, which amounts to approximately 12.6 million workers. Notably, and differently from other extant measures, our approach allows to identify not a point value but rather an interval of exposure which enables to understand how labour-substitutability unevenly hampers the labour force.

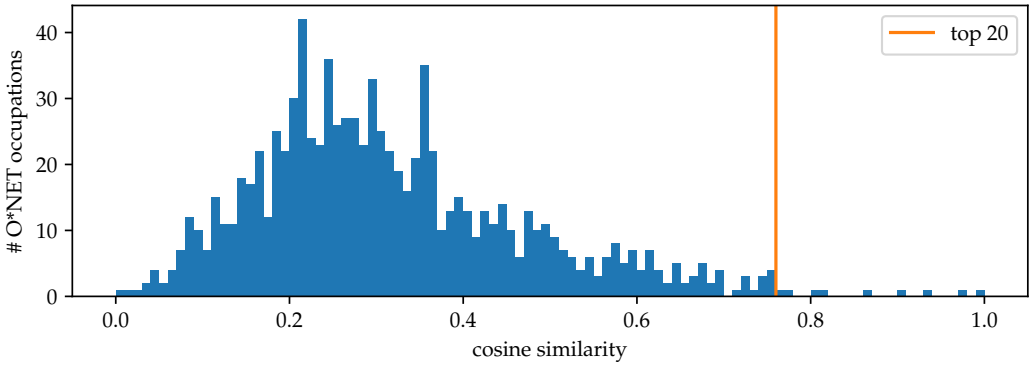


Figure 6: Distribution of cosine similarity with respect to 8-digit occupations.

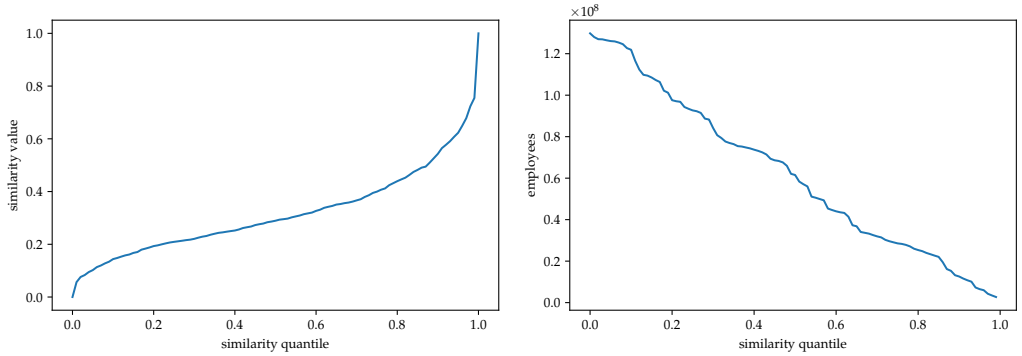


Figure 7: Quantile function of the similarity distribution for 8-digit occupations (left) and number of replaceable employees by quantile (right).

| Rank | Code | Description | CS |
|---|---|---|---|
| 1 | 14587 | Load materials and products into machines and equipment, or onto conveyors, using hand tools and moving devices | 1.0 |
| 2 | 3202 | Move levers or controls that operate lifting devices, such as forklifts, lift beams with swivel-hooks, hoists, or elevating platforms, to load, unload, transport, or stack material | 0.96 |
| 3 | 3203 | Position lifting devices under, over, or around loaded pallets, skids, or boxes and secure material or products for transport to designated areas | 0.9 |
| 4 | 17928 | Lift and move loads, using cranes, hoists, and rigging, to install or repair hydroelectric system equipment or infrastructure | 0.89 |
| 5 | 15266 | Manually or mechanically load or unload materials from pallets, skids, platforms, cars, lifting devices, or other transport vehicles | 0.88 |
| 6 | 14584 | Remove materials and products from machines and equipment, and place them in boxes, trucks or conveyors, using hand tools and moving devices | 0.86 |
| 7 | 11839 | Transport machine parts, tools, equipment, and other material between work areas and storage, using cranes, hoists, or dollies | 0.85 |
| 8 | 3217 | Load materials and products into package processing equipment | 0.85 |
| 9 | 12805 | Operate conveyors and equipment to transfer grain or other materials from transportation vehicles | 0.85 |
| 10 | 12323 | Communicate with systems operators to regulate and coordinate line voltages and transmission loads and frequencies | 0.84 |
| 11 | 12798 | Operate industrial trucks, tractors, loaders, and other equipment to transport materials to and from transportation vehicles and loading docks, and to store and retrieve materials in warehouses | 0.83 |
| 12 | 20387 | Optimize photonic process parameters by making prototype or production devices | 0.83 |
| 13 | 17496 | Provide information about community health and social resources | 0.83 |
| 14 | 13705 | Unload materials, devices, and machine parts, using hand tools | 0.8 |
| 15 | 10757 | Load, unload, or adjust materials or products on conveyors by hand, by using lifts, hoists, and scoops, or by opening gates, chutes, or hoppers | 0.8 |

Table 2: Top 15 tasks by (rescaled) similarity.

| Rank | Code | Title | CS |
|---|---|---|---|
| 1 | 53-7051.00 | Industrial Truck and Tractor Operators | 1.0 |
| 2 | 49-9043.00 | Maintenance Workers, Machinery | 0.97 |
| 3 | 53-7063.00 | Machine Feeders and Offbearers | 0.94 |
| 4 | 53-7064.00 | Packers and Packagers, Hand | 0.91 |
| 5 | 49-2091.00 | Avionics Technicians | 0.87 |
| 6 | 51-9111.00 | Packaging and Filling Machine Operators and Tenders | 0.81 |
| 7 | 49-3041.00 | Farm Equipment Mechanics and Service Technicians | 0.81 |
| 8 | 49-3092.00 | Recreational Vehicle Service Technicians | 0.78 |
| 9 | 49-3042.00 | Mobile Heavy Equipment Mechanics, Except Engines | 0.77 |
| 10 | 47-2111.00 | Electricians | 0.76 |
| 11 | 49-9098.00 | Helpers–Installation, Maintenance, and Repair Workers | 0.75 |
| 12 | 49-9041.00 | Industrial Machinery Mechanics | 0.75 |
| 13 | 51-9082.00 | Medical Appliance Technicians | 0.75 |
| 14 | 47-3011.00 | Helpers–Brickmasons, Blockmasons, Stonemasons, and Tile and Marble Setters | 0.75 |
| 15 | 51-9191.00 | Adhesive Bonding Machine Operators and Tenders | 0.75 |
| 16 | 51-9023.00 | Mixing and Blending Machine Setters, Operators, and Tenders | 0.74 |
| 17 | 13-1032.00 | Insurance Appraisers, Auto Damage | 0.73 |
| 18 | 51-4111.00 | Tool and Die Makers | 0.73 |
| 19 | 49-9081.00 | Wind Turbine Service Technicians | 0.72 |
| 20 | 51-8013.04 | Hydroelectric Plant Technicians | 0.72 |

Table 3: Top 20 occupations by (rescaled) similarity.

### 5.2 Industry and labour market penetration

Occupations are distributed across industries, and therefore identifying the most and least affected ones is crucial for any potential policy intervention. Table 4 shows the relevance of occupational exposure to robotic LS technologies in each NAICS 2-digit sector by weighting the cosine similarity by percentage of occupation membership to each sector. The measure, which takes value 1 for the most and value 0 for the least exposed industry (in relative terms), depicts the manufacturing sector as the most exposed to automation. Not only manufacturing is leading the ranking, but all the other sectors follow after a dramatic drop. Indeed, the industry ranking reflects the aggregation of scattered occupations entailing manual abilities and handling, as shown above, which however are largely concentrated in manufacturing. Moreover, manufacturing as an industry collects all those activities related to logistics and warehouse which are still currently under the parent manufacturing companies, while third-party logistics, doing logistics activities as a service, rank seventh across industries.

On the whole, robotic LS technologies will further deepen the long-run ongoing de-industrialisation of the US economy (cf. Section 3). However it is remarkable the high ranking of healthcare, social assistance and education, the most human-oriented industries. Such a result is a warning, already detected in Montobbio et al. (2022), about the direction of cutting-edge innovative efforts diverted towards industries where, at least in principle, the human-based component should be preferable. Similarly, public administration ranks in the top-five most exposed sectors to substitution and this signals the ability of our measure to comprise not only automation *per se* but also advanced digitalisation processes entailed by administrative services. Notably, *"Management of Companies and Enterprises"*, although recording a contraction in occupational employment shares in the last twenty years, presents the lowest similarity score.

The last battery of results is shown in Figs. 8 and 9, presenting a non-parametric LOWESS estimation (locally weighted scatterplot smoothing), as in Acemoglu and Autor (2011) and Webb (2020), of the relationship between the cosine similarity measure, employment, and wages both in level in 2019 and in growth rate in 1999-2019. A neat negative, almost monotonically decreasing relationship emerges in all four considered variables. Starting with employment levels, LOWESS estimates confirm the previous evidence, showing low occurrence of high similarity measures for the majority of employees. However, such negative relationship also emerges when the similarity measure is compared against employment growth, revealing that shrinking occupations in the last two decades have been also those most exposed to robotic LS technologies. Such evidence confirms that, among

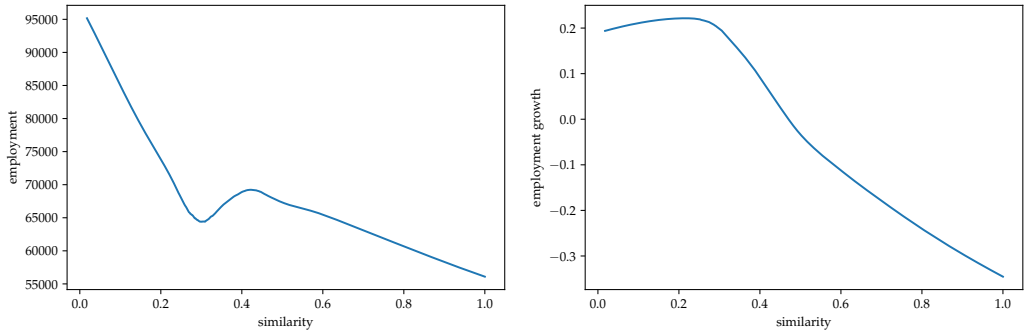


Figure 8: Similarity and employment level in 2019 (left) and 1999-2019 employment growth (right). Robust LOWESS estimates of the underlying scatter plots (bandwidth = 0.8).

other potential sources determining the displacement of some occupations, substitutionary technical change related to automation might have played a role.

Towards which labour force segments are such innovative efforts directed? According to the LOWESS estimates, the relationship in terms of wage level and growth, again almost monotonically decreasing, signals that the most exposed occupations to LS technologies are the least-paid and recording the lowest wage growth. In other words, robotic LS technologies and their underlying patents are more directed towards substituting the cheapest segments of the labour force.

Anecdotal evidence suggests a high incidence of highly automatised production processes in already quite standardised workplaces (Ford, 2015); related, a large majority of case studies on Industry 4.0 questions the revolutionary content of the latest technological wave and highlights the patterns of continuity with ICT (Cetrulo and Nuvolari, 2019; Cirillo et al., 2021; Dosi and Virgillito, 2019; Krzywdzinski, 2020; Santarelli et al., 2021). High innovative efforts to automate cheap labour are what Acemoglu and Restrepo (2020b) define *"so-so"* technologies. Far from judging labour from its remuneration, it is evident from our analysis that gains from automation in terms of productivity are searched in the knowledge and technological space allowing for incremental upgrading of already automated processes and substituting the labour force therein involved.

### 5.3 Geographical penetration

The United States are very much differentiated in terms of productive specialisation and ensuing occupational composition. Understanding the different geographical penetration of robotic LS technologies across states is useful and informative, both as a validation exercise and as a tool to perform targeted policy actions.

| Rank | NAICS | CS |
|---|---|---|
| 1 | Manufacturing | 1.0 |
| 2 | Health care and social assistance | 0.39 |
| 3 | Education services | 0.33 |
| 4 | Construction | 0.30 |
| 5 | Public administration | 0.21 |
| 6 | Other services, except public administration | 0.18 |
| 7 | Transportation and warehousing | 0.17 |
| 8 | Retail trade | 0.16 |
| 9 | Professional, Scientific and Technical Services | 0.11 |
| 10 | Utilities | 0.10 |
| 11 | Administrative and support and waste management and remediation services | 0.09 |
| 12 | Information | 0.09 |
| 13 | Arts, entertainment, and recreation | 0.07 |
| 14 | Accommodation and food services | 0.07 |
| 15 | Wholesale trade | 0.07 |
| 16 | Mining | 0.06 |
| 17 | Finance and insurance | 0.05 |
| 18 | Agriculture, forestry, fishing and hunting | 0.04 |
| 19 | Real estate and rental and leasing | 0.02 |
| 20 | Management of Companies and Enterprises | 0.00 |

Table 4: Relevance of exposed occupations to NAICS 2-digit sectors, obtained as a weighted average of similarity and occupation membership to the underlying sector, rescaled between 0 and 1.

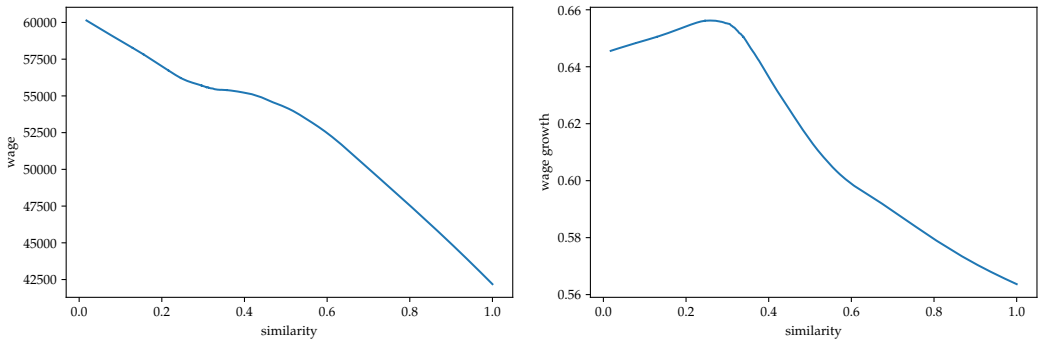


Figure 9: Similarity and wage level in 2019 (left) and 1999-2019 wage growth (right). Robust LOWESS estimates of the underlying scatter plots (bandwidth $= 0.8$).

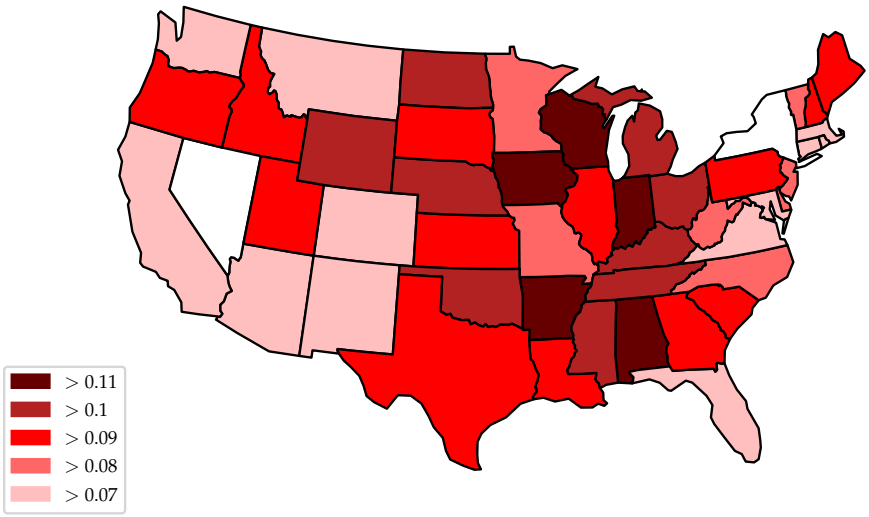


Figure 10: Disaggregation by state. Employment shares of most exposed occupations (top-decile) to LS technologies. Continental US.

Fig. 10 shows the state-level disaggregation of the top-exposed occupations (top-decile of the distribution in Fig. 6). The heat-map presents five colour shades, going from the more exposed states to LS technologies (dark brown), to less exposed states (light pink). The range of variation, considering the 8.6% share for the US as a whole, goes from 4% to 11%. In the picture, states presenting a share below 7% are coloured in white.

Going from the most to the least exposed states, according to our results, the so-called Rust Belt area (i.e. the region surrounding the Great Lakes) is populated by states characterised by darker reds, namely Wisconsin, Indiana, Michigan, Illinois. Other dark red states are Alabama, Arkansas, Mississippi, Louisiana, located in the Southern East, where African-Americans are largely concentrated. Texas, with a prevalent fraction of Hispanic communities, presents 9% of the occupational categories exposed to LS technologies. At the opposite, all states on the two coasts, representing the cradle of high-tech companies and of high-specialised service activities are coloured in light red, like Washington, California, Florida, Virginia. Notably, states like New York and D.C. are indicated in white being the share of occupations exposed to LS technologies below 7%.

# 6 Conclusions

This paper represents one of the first attempts at building a *direct* measure of occupational exposure to labour-saving technologies. We encompass such an objective by making use of natural language processing (NLP) techniques. First, we leverage on the information retrieved by Montobbio et al. (2022), allowing to identify explicitly labour-saving robotic patent applications, collected in the 2008-2018 period. After identifying robotic and LS robotic patents, the underlying 4-digit CPC definitions are employed in order to detect functions and operations performed by technological artefacts which are more directed to substitute the labour input. The measure allows to obtain fine-grained information on tasks and occupations according to their similarity ranking. In addition, performing a match with the occupational and wage statistics, we offer evidence on the relationship between our exposure measure, employment and wage levels, and their changes in the last twenty years. Industry and geographical penetrations are studied as well.

According to our results, the occupations which are more exposed to labour-saving automation are those performing activities related to manual dexterity, manipulation, loading of objects into machines and equipment, lift and load moving. These tasks are mainly characterising occupations as industrial truck and tractor operators, packaging and filling machine operators and tenders, tool and die makers, but also medical appliance technicians. When decomposing the information by industry, we do find that manufacturing is the most exposed one, followed by healthcare, social assistance, and education services. In addition, although manufacturing is the most exposed sector, many identified tasks regard logistics activities.

LOWESS estimates of the relationship between occupational exposure and labour market variables confirm that robotic labour-saving innovations target low paid occupations, experiencing the lowest wage increases in the last twenty years. In addition, such low paid occupations are also shrinking in terms of workforce. Although we cannot encompass for the entire set of potential confounding factors affecting such relationship, we are able to find striking and clear-cut patterns of innovative efforts directed toward the cheapest and most vulnerable segments of the working population.

Exposure to LS automation and employment dynamics do not always go hand in hand. Take the 2-digit occupation *"Transportation and Material Moving"*, very much exposed to labour-saving technologies, which has however experienced a positive employment growth (cf. Fig. 1). Whenever occupations record a positive growth, notwithstanding their high exposure rate, this signals that employment dynamics is driven by other sources, primarily demand, which might clearly counter-

balance the potential labour-saving traits of advancement in robotics. Notably, the aggregation at the industry level highlights a deepening of the de-industrialisation trend of the US economy, since manufacturing is by far the most exposed sector. However, social care and assistance services, as well as education, turn out to be quite high in the exposure ranking. Therefore, not only low-paid manufacturing and logistics workers are exposed, but also low-paid service workers. On the contrary, managerial occupations, although reducing in employment shares, present the lowest degree of similarity.

A companion result of our study is that high similarity is a quite rare event: the CPC-task cosine similarity matrix is sparse and high values of similarity are more the exception than the norm. This finding corroborates our procedure, whose results are not inflated but rather conservatively underestimated, given our very cautious identification of robotic labour-saving patents (cf. Section 4.1). As a consequence, when considering the cumulative fraction of potentially replaceable occupations, the top-decile of the similarity distribution involves 8.6% or approximately 12.6 million employees.

Strengths of our approach are, first, the construction of a direct measure of proximity by means of an objective procedure, not resorting to subjective and mutable expert judgments, or alternatively crowdsourcing. Second, the generality, being the measure constructed on the *entire* set of CPC codes, and only in a second step using a weighting procedure to account for labour-saving technologies. This means that we obtain a similarity measure for the entire technological (CPC) and occupational (O*NET tasks) spectra. Finally, the non linear-nature of the quantile threshold and the sparsity nature of the matrix comfort both in terms of reliability and in terms of labour market prospects.

A first limitation of the present study is that it gives account uniquely of robotic labour-saving innovations, while labour-saving innovations encompass both other applications of AI technologies such as technological change embodied in machineries and tools distinct from robots. A second limitation is that we are not able to track adopters of these technologies and we do not know the exact number of workers, in terms of intensive rather than extensive margin, each machine embedding a labour-saving technology is able to replace. This means that, potentially, if adopters are widespread and the number of their labour units is high, the occupational losses might be much higher than predicted in this work.

Potential extensions of our study entail, first, the LS identification of other technologies beyond strict robotics ones, such as AI or standard ICT. Second, our measure can be adopted to other labour markets, beyond the US. Third, an application of our indicator at the firm-level constitutes a quite promising avenue of research,

in order to pinpoint the establishments and plants more exposed to labour-saving efforts.

## Acknowledgements

The authors wish to thank participants at the 2021 Annual Conference of the Italian Economic Association, the 2021 Annual Conference of the European Association for Evolutionary Political Economy, the 2021 International Schumpeter Society Conference, and the 2021 Conference of the European Network on the Economics of the Firm for helpful comments and insightful suggestions at various stages of this work. Fabio Montobbio, Jacopo Staccioli, and Marco Vivarelli acknowledge support by the Italian Ministero dell'Istruzione, dell'Università e della Ricerca, PRIN-2017 project 201799ZJSN: "Technological change, industry evolution and employment dynamics" (principal investigator: Marco Vivarelli). Maria Enrica Virgillito acknowledges support from European Union's Horizon 2020 research and innovation programme under grant agreement No. 822781 "GROWINPRO – Growth Welfare Innovation Productivity".

## References

Acemoglu, Daron and David H. Autor (2011). 'Skills, Tasks and Technologies: Implications for Employment and Earnings'. In: *Handbook of Labor Economics*. Ed. by David Card and Orley Ashenfelter. Vol. 4B. Elsevier. Chap. 12, pp. 1043–1171. DOI: 10.1016/S0169-7218(11)02410-5.

Acemoglu, Daron, David H. Autor, Jonathon Hazell and Pascual Restrepo (2020a). *AI and jobs: Evidence from online vacancies*. NBER Working Paper 28257. DOI: 10.3386/w28257.

Acemoglu, Daron, Claire Lelarge and Pascual Restrepo (2020b). 'Competing with robots: Firm-level evidence from france'. In: *AEA Papers and Proceedings*. Vol. 110, pp. 383–88.

Acemoglu, Daron and Pascual Restrepo (2018). 'The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment'. *American Economic Review* 108(6), pp. 1488–1542. DOI: 10.1257/aer.20160696.

Acemoglu, Daron and Pascual Restrepo (2019). 'Automation and New Tasks: How Technology Displaces and Reinstates Labor'. *Journal of Economic Perspectives* 33(2), pp. 3–30. DOI: 10.1257/jep.33.2.3.

Acemoglu, Daron and Pascual Restrepo (2020a). 'Robots and Jobs: Evidence from US Labor Markets'. *Journal of Political Economy* 128(6), pp. 2188–2244. DOI: 10.1086/705716.

Acemoglu, Daron and Pascual Restrepo (2020b). 'The wrong kind of AI? Artificial intelligence and the future of labour demand'. *Cambridge Journal of Regions, Economy and Society* 13(1), pp. 25–35. DOI: 10.1093/cjres/rsz022.

Aggarwal, Charu C. (2018). *Machine Learning for Text*. Springer.

Arntz, Melanie, Terry Gregory and Ulrich Zierahn (2016). *The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis*. OECD Social, Employment and Migration Working Papers No. 189, OECD Publishing, Paris. DOI: 10.1787/5jlz9h56dvq7-en.

Autor, David H. and David Dorn (2013). 'The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market'. *American Economic Review* 103(5), pp. 1553–1597. DOI: 10.1257/aer.103.5.1553.

Autor, David, Anna Salomons and Bryan Seegmiller (2020). *New Frontiers: The Origins and Content of New Work, 1940–2018*. Tech. rep. MIT Mimeo.

Barbieri, Laura, Mariacristina Piva and Marco Vivarelli (2018). 'R&D, embodied technological change, and employment: evidence from Italian microdata'. *Industrial and Corporate Change* 28(1), pp. 203–218. DOI: 10.1093/icc/dty001.

Brynjolfsson, Erik and Andrew McAfee (2012). *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Digital Frontier Press.

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

Brynjolfsson, Erik, Tom Mitchell and Daniel Rock (2018). 'What can machines learn, and what does it mean for occupations and the economy?' In: *AEA Papers and Proceedings*. Vol. 108, pp. 43–47. DOI: 10.1257/pandp.20181019.

Calvino, Flavio and M. Enrica Virgillito (2018). 'The innovation employment nexus: A critical survey of theory and empirics'. *Journal of Economic Surveys* 32(1), pp. 83–117. DOI: 10.1111/joes.12190.

Cetrulo, Armanda and Alessandro Nuvolari (2019). 'Industry 4.0: revolution or hype? Reassessing recent technological trends and their impact on labour'. *Journal of Industrial and Business Economics* 46(3), pp. 391–402. DOI: 10.1007/s40812-019-00132-y.

Chiacchio, Francesco, Georgios Petropoulos and David Pichler (2018). 'The impact of industrial robots on EU employment and wages: A local labour market approach'. Bruegel Working Paper, No. 2018/02. URL: https://www.econstor.eu/handle/10419/207001.

Cirillo, Valeria, Matteo Rinaldini, Jacopo Staccioli and M. Enrica Virgillito (2021). 'Technology vs. workers: the case of Italy's Industry 4.0 factories'. *Structural Change and Economic Dynamics* 56, pp. 166–183. DOI: 10.1016/j.strueco.2020.09.007.

Clark, John, Christopher Freeman and Luc Soete (1981). 'Long waves and technological developments in the 20th century'. *Konjunktur, Krise, Gesellschaft* 25(2), pp. 132–169.

Dauth, Wolfgang, Sebastian Findeisen, Jens Südekum and Nicole Woessner (2017). 'German Robots – The Impact of Industrial Robots on Workers'. CEPR Discussion Paper No. DP12306, Available at SSRN. URL: https://ssrn.com/abstract=3039031.

Domini, Giacomo, Marco Grazzi, Daniele Moschella and Tania Treibich (2020). 'Threats and opportunities in the digital era: Automation spikes and employment dynamics'. *Research Policy*. In press. DOI: 10.1016/j.respol.2020.104137.

Dosi, Giovanni and Pierre Mohnen (2019). 'Innovation and employment: an introduction'. *Industrial and Corporate Change* 28(1), pp. 45–49. DOI: 10.1093/icc/dty064.

Dosi, Giovanni, Mariacristina Piva, Maria Enrica Virgillito and Marco Vivarelli (2021). 'Embodied and Disembodied Technological Change: The Sectoral Patterns of Job-Creation and Job-Destruction'. *Research Policy* 50(4), p. 104199. DOI: 10.1016/j.respol.2021.104199.

Dosi, Giovanni and Maria Enrica Virgillito (2019). 'Whither the evolution of the contemporary social fabric? New technologies and old socio-economic trends'. *International Labour Review* 158(4), pp. 593–625. DOI: 10.1111/ilr.12145.

Felten, Edward W, Manav Raj and Robert Seamans (2018). 'A method to link advances in artificial intelligence to occupational abilities'. In: *AEA Papers and Proceedings*. Vol. 108, pp. 54–57. DOI: 10.1257/pandp.20181021.

Felten, Edward, Manav Raj and Robert Seamans (2021). 'Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses'. *Strategic Management Journal* [forthcoming]. DOI: 10.1002/smj.3286.

Ford, Martin (2015). *The Rise of the Robots. Technology and the Threat of Mass Unemployment*. Basic Books.

Freeman, Christopher and Luc Soete (1987). *Technical change and full employment*. B. Blackwell.

Frey, Carl Benedikt and Michael A. Osborne (2017). 'The future of employment: How susceptible are jobs to computerisation?' *Technological Forecasting and Social Change* 114, pp. 254–280. DOI: 10.1016/j.techfore.2016.08.019.

Graetz, Georg and Guy Michaels (2018). 'Robots at Work'. *Review of Economics and Statistics* 100(5), pp. 753–768. DOI: 10.1162/rest_a_00754.

Handel, Michael J (2016). 'The O* NET content model: strengths and limitations'. *Journal for Labour Market Research* 49(2), pp. 157–176. DOI: 10.1007/s12651-016-0199-8.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru and Matt Taddy (2018). *Measuring technological innovation over the long run*. Tech. rep. National Bureau of Economic Research.

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

Kogan, Leonid, Dimitris Papanikolaou, Lawrence D.W. Schmidt and Bryan Seegmiller (2021). *Technology-Skill Complementarity and Labor Displacement: Evidence from Linking Two Centuries of Patents with Occupations*. Mimeo. URL: https://www.bryanseegmiller.com/files/Draft_v2021-1017.pdf.

Krzywdzinski, Martin (2020). *Automation, Digitalization, and Changes in Occupational Structures in the Automobile Industry in Germany, the United States, and Japan: A Brief History from the Early 1990s Until 2018*. (Weizenbaum Series #10) Berlin: Weizenbaum Institute for the Networked Society – The German Internet Institute. DOI: 10.34669/wi.ws/10.

Meindl, Benjamin, Morgan R Frank and Joana Mendonça (2021). 'Exposure of occupations to technologies of the fourth industrial revolution'. *arXiv preprint arXiv:2110.13317*.

Mondolo, Jasmine (2021). 'The composite link between technological change and employment: A survey of the literature'. *Journal of Economic Surveys*. [forthcoming]. DOI: 10.1111/joes.12469.

Montobbio, Fabio, Jacopo Staccioli, Maria Enrica Virgillito and Marco Vivarelli (2022). 'Robots and the origin of their labour-saving impact'. *Technological Forecasting and Social Change* 174, 121122. DOI: 10.1016/j.techfore.2021.121122.

Nedelkoska, Ljubica and Glenda Quintini (2018). *Automation, skills use and training*. OECD Social, Employment and Migration Working Papers, No. 202, OECD Publishing, Paris. DOI: 10.1787/2e2f4eea-en.

Pellegrino, Gabriele, Mariacristina Piva and Marco Vivarelli (2019). 'Beyond R&D: the role of embodied technological change in affecting employment'. *Journal of Evolutionary Economics* 29(4), pp. 1151–1171. DOI: 10.1007/s00191-019-00635-w.

Piva, Mariacristina and Marco Vivarelli (2018). 'Technological change and employment: is Europe ready for the challenge?' *Eurasian Business Review* 8(1), pp. 13–32. DOI: 10.1007/s40821-017-0100-x.

Santarelli, Enrico, Jacopo Staccioli and Marco Vivarelli (2021). 'Robots, AI, and Related Technologies: A Mapping of the New Knowledge Base'. LEM Working Paper Series n. 01/2021. URL: http://www.lem.sssup.it/WPLem/2021-01.html.

Simonetti, Roberto, Karl Taylor and Marco Vivarelli (2000). 'Modelling the employment impact of innovation'. In: *The Employment Impact of Innovation*. Ed. by Mario Pianta and Marco Vivarelli. Routledge. Chap. 3, pp. 26–43.

Staccioli, Jacopo and Maria Enrica Virgillito (2021). 'Back to the past: the historical roots of labor-saving automation'. *Eurasian Business Review* 11(1), pp. 27–57.

Van Roy, Vincent, Dániel Vértesy and Marco Vivarelli (2018). 'Technology and employment: Mass unemployment or job creation? Empirical evidence from European patenting firms'. *Research Policy* 47(9), pp. 1762–1776. DOI: 10.1016/j.respol.2018.06.008.

Vivarelli, Marco (1995). *The Economics of Technology and Employment: Theory and Empirical Evidence*. Edward Elgar Pub.

Vivarelli, Marco (2015). 'Innovation and employment'. IZA World of Labor. URL: https://wol.iza.org/articles/innovation-and-employment/long.

Webb, Michael (2020). 'The Impact of Artificial Intelligence on the Labor Market'. Available at SSRN. URL: https://ssrn.com/abstract=3482150.

# Appendix A

In this Appendix we briefly summarise the methodological steps adopted in Montobbio et al. (2022).

**Step 1 – Identification of robotic patents** The analysis starts with the entire set of 3,557,435 patent applications published by the USPTO between 1st January 2009 and 31st December 2018. Robotic patents are pinpointed therein according to two distinct criteria, one based on the patent classification codes specified within applications, the other based on textual keyword search. A patent is deemed 'robotic' if it obeys at least one of the criteria. In particular, a robotic patent according to the first criterion (dubbed `'CPC'`) must be assigned by patent examiners at least one of a set of 174 full-digit CPC codes which reflect former US. Patent Classification (USPC) class 901 (*"Robots"*). Likewise, a robotic patent according to the second criterion (dubbed `'K10'`) must contain the word `'robot'` in its full-text at least 10 times, including derivational and inflectional affixes. The first criterion identifies 10,929 robotic patents, while the second criterion identifies another 18,860 (after discarding robotic patents according to the first criterion). The two criteria single out a total of 29,789 robotic patents, i.e. approximately 0.84% of the original (universe) population. Their evolution over time is shown in Fig. 3 (Section 3).

**Step 2 – Identification of labour-saving patents** LS patents constitute a subset of robotic patents, identified by a multiple *word* co-occurrence query at the *sentence* level. In particular, a patent is deemed LS (after an additional manual validation step) if its full-text contains at least one sentence in which the verbal predicate, direct object, and object attribute belong to the following lists:

$$\underbrace{\begin{bmatrix} \text{'reduc'} \\ \text{'replac'} \\ \text{'elimin'} \\ \text{'save'} \\ \text{'lower'} \\ \text{'substitut'} \\ \text{'autom'} \end{bmatrix}}_{\text{verbal predicate}} \times \underbrace{\begin{bmatrix} \text{'labor'} \\ \text{'worker'} \\ \text{'human'} \\ \text{'employe'} \\ \text{'manpow'} \\ \text{'job'} \end{bmatrix}}_{\text{direct object}} \times \underbrace{\begin{bmatrix} \text{'cost'} \\ \text{'expenditure'} \\ \text{'expens'} \\ \text{'hour'} \\ \text{'intens'} \\ \text{'task'} \\ \text{'time'} \\ \text{'skill'} \end{bmatrix}}_{\text{object attribute}}.$$

In total, 1,276 LS patents are found (approximately 4.3% of all robotic patents), of which 461 ($\approx$ 36.1%) belong to the `CPC` group and 815 ($\approx$ 63.9%) belong to the `K10` group.

**Step 3 – Identification of the underlying technology** Table 5 reports the top ten couples and triplets of 3-digit CPC codes recurring in our patents, ranked by frequency, whose description is presented in Table 6. These account for a sizeable chunk of the identified robotics and LS patents, respectively. The focus on couples and triplets of co-occurring classification codes allows us to better grasp the underlying technological complementarities. Indeed, our patents, both robotic and LS, are characterised by a strong recurrent pattern of coupling between technologies belonging to automation (such as codes `B25`, `B65`) and codes related to computing and information processing (such as those belonging to the `G` codes). Additionally, against a potential bias towards industrial robots, we see the emergence of CPC code `A61`, explicitly related to healthcare.

To further characterise the extent to which the selected patents are complementary to the field of AI, i.e. embed traits of "intelligent automation", we check their degree of overlap with AI patents identified by Santarelli et al. (2021), who use an analogous methodology which leverages CPC codes obtained through a statistical concordance table with USPC class 706, *"Artificial intelligence"*. We find that while 10.5% of robotic patents (3,140 units) are also classifiable as AI, when restricting the analysis to LS patents this percentage increases to 23.5% (300 units). This evidence supports the notion of intelligent automation or intelligent robots or automated production processes, particularly when referring to LS patents.

## Appendix B

In order to test the robustness of our procedure, we replicate the text similarity exercise described in Section 4 using LS patents full-texts, rather than CPC code definitions. In principle, one might expect the underlying content of the patent to be more informative than CPC definitions when it comes to actual labour-saving efforts and the eventual human functions substituted.

Tables 7 and 8 present the results of the top-task and top-occupation matching by similarity scores. Interestingly, the identified tasks and occupations are completely different from the original exercise. Emerging tasks are *"Build or assemble robotic devices or systems"*; *"Set up and operate computer-controlled machines or robots to perform one or more machine functions on metal or plastic workpieces"*; *"Build, configure, or test robots or robotic applications"*; *"Conduct research on robotic technology to create new robotic systems or system capabilities"*; *"Provide technical support for robotic systems"*. These tasks are clearly the ones labour-complementing, i.e. required to develop and manufacture the new robotic artefacts. Occupations more exposed to labour-complementarity are indeed *"Robotics Engineers"*; *"Robotics Technicians"*;

| Robotic patents | | | LS patents | | |
|---|---|---|---|---|---|
| **CPC couples** | **Count** | **Frequency** | **CPC couples** | **Count** | **Frequency** |
| B25, G05 | 4,169 | 0.073 | B25, G05 | 272 | 0.068 |
| G05, G06 | 1,879 | 0.033 | G05, G06 | 193 | 0.048 |
| B25, G05 | 4,169 | 0.073 | G06, H04 | 131 | 0.033 |
| G06, H04 | 1,476 | 0.026 | B25, G06 | 125 | 0.031 |
| A61, B25 | 1,248 | 0.022 | B25, B65 | 93 | 0.023 |
| G01, G05 | 1,054 | 0.018 | B65, G06 | 87 | 0.022 |
| B25, G01 | 894 | 0.016 | B65, G05 | 76 | 0.019 |
| G01, G06 | 796 | 0.014 | G05, H04 | 76 | 0.019 |
| G05, H04 | 745 | 0.013 | B23, B25 | 59 | 0.015 |
| A61, G06 | 732 | 0.013 | B25, H04 | 51 | 0.013 |
| **CPC triplets** | **Count** | **Frequency** | **CPC triplets** | **Count** | **Frequency** |
| B25, G05, G06 | 1,015 | 0.024 | B25, G05, G06 | 98 | 0.025 |
| G05, G06, H04 | 474 | 0.011 | G05, G06, H04 | 67 | 0.017 |
| A61, B25, G05 | 405 | 0.01 | B25, B65, G05 | 51 | 0.013 |
| B25, G01, G05 | 401 | 0.01 | B25, G06, H04 | 46 | 0.012 |
| B25, G05, H04 | 371 | 0.009 | B65, G05, G06 | 43 | 0.011 |
| B25, G06, H04 | 348 | 0.008 | B25, G05, H04 | 42 | 0.011 |
| G01, G05, G06 | 335 | 0.008 | A01, B25, G05 | 36 | 0.009 |
| G01, G06, H04 | 288 | 0.007 | A01, G05, G06 | 33 | 0.009 |
| B23, B25, G05 | 246 | 0.006 | B25, B65, G06 | 32 | 0.008 |
| A61, G06, G16 | 244 | 0.006 | A01, B25, G06 | 32 | 0.008 |

Table 5: Top ten couples and triplets of 3-digit CPC codes assigned to robotic (left half) and LS patents (right half), respectively. Source: Montobbio et al. (2022, Tab. 1).

| CPC | Definition |
|---|---|
| A01 | Agriculture; Forestry; Animal husbandry; Hunting; Trapping; Fishing |
| A61 | Medical or veterinary science; Hygiene |
| B23 | Machine tools; Metal-working not otherwise provided for |
| B25 | Hand tools; Portable power-driven tools; Manipulators |
| B65 | Conveying; Packing, Storing; Handling thin or filamentary material |
| G01 | Measuring; Testing |
| G05 | Controlling; Regulating |
| G06 | Computing; Calculating; Counting |
| G16 | Information and communication technology (ICT) specially adapted for specific fields |
| H04 | Electric communication technique |

Table 6: Main 3-digit CPC codes definitions. Source: Montobbio et al. (2022, Tab. 2).

*"Computer Systems Engineers/Architects"*; *"Data Warehousing Specialists"*; *"Network and Computer Systems Administrators"*.

Notably, the similarity measure by occupations presents a drop after the first two most-exposed occupations onwards. In addition, the elicited occupations reveal who are those workers programming and creating LS technologies: they belong to the upper-echelon of the occupational categories, are well paid, and growing in number during the last two decades. Indeed, top-complementary occupations tend to belong to *"Computer and Mathematical Occupations"* (15) and *"Architecture and Engineering Occupations"* (17).

| Rank | Code | Description | CS |
|---|---|---|---|
| 1 | 16596 | Build or assemble robotic devices or systems | 1.0 |
| 2 | 11944 | Set up and operate computer-controlled machines or robots to perform one or more machine functions on metal or plastic workpieces | 0.98 |
| 3 | 21057 | Build, configure, or test robots or robotic applications | 0.97 |
| 4 | 16523 | Conduct research on robotic technology to create new robotic systems or system capabilities | 0.93 |
| 5 | 16511 | Provide technical support for robotic systems | 0.91 |
| 6 | 16587 | Assist engineers in the design, configuration, or application of robotic systems | 0.86 |
| 7 | 16525 | Conduct research into the feasibility, design, operation, or performance of robotic mechanisms, components, or systems, such as planetary rovers, multiple mobile robots, reconfigurable robots, or man-machine interactions | 0.84 |
| 8 | 16593 | Install, program, or repair programmable controllers, robot controllers, end-of-arm tools, or conveyors | 0.81 |
| 9 | 16584 | Modify computer-controlled robot movements | 0.8 |
| 10 | 16579 | Maintain service records of robotic equipment or automated production systems | 0.8 |
| 11 | 20262 | Plan special events, parties, or meetings, which may include booking musicians or celebrities | 0.82 |
| 12 | 14861 | Inquire about pesticides or chemicals to which animals may have been exposed | 0.79 |
| 13 | 16075 | Implement controls to provide security for operating systems, software, and data | 0.79 |
| 14 | 23748 | Prepare and submit reports that may include the number of passengers or trips, hours worked, mileage, fuel consumption, or fares received | 0.79 |
| 15 | 1277 | Perform systems analysis and programming tasks to maintain and control the use of computer systems software as a systems programmer | 0.79 |
| 16 | 2463 | Develop networks of attorneys, mortgage lenders, and contractors to whom clients may be referred | 0.78 |
| 17 | 246 | Arrange for medical, psychiatric, and other tests that may disclose causes of difficulties and indicate remedial measures | 0.77 |
| 18 | 11338 | Transport mail from one work station to another | 0.77 |
| 19 | 18280 | Install, calibrate, or maintain sensors, mechanical controls, GPS-based vehicle guidance systems, or computer settings | 0.77 |
| 20 | 16434 | Calibrate vehicle systems, including control algorithms or other software systems | 0.77 |

Table 7: Top 20 tasks by (rescaled) similarity (based on LS patents full-texts).

| Rank | Code | Title | CS |
|---|---|---|---|
| 1 | 17-2199.08 | Robotics Engineers | 1.0 |
| 2 | 17-3024.01 | Robotics Technicians | 0.96 |
| 3 | 47-2231.00 | Solar Photovoltaic Installers | 0.49 |
| 4 | 17-2072.01 | Radio Frequency Identification Device Specialists | 0.46 |
| 5 | 15-1299.08 | Computer Systems Engineers/Architects | 0.45 |
| 6 | 15-1299.02 | Geographic Information Systems Technologists and Technicians | 0.42 |
| 7 | 51-9161.00 | Computer Numerically Controlled Tool Operators | 0.41 |
| 8 | 17-2199.11 | Solar Energy Systems Engineers | 0.4 |
| 9 | 49-2091.00 | Avionics Technicians | 0.39 |
| 10 | 15-1243.01 | Data Warehousing Specialists | 0.38 |
| 11 | 17-1022.01 | Geodetic Surveyors | 0.38 |
| 12 | 15-1244.00 | Network and Computer Systems Administrators | 0.38 |
| 13 | 17-2061.00 | Computer Hardware Engineers | 0.37 |
| 14 | 15-1299.03 | Document Management Specialists | 0.37 |
| 15 | 15-1211.00 | Computer Systems Analysts | 0.36 |
| 16 | 51-4034.00 | Lathe and Turning Machine Tool Setters, Operators, and Tenders, Metal and Plastic | 0.36 |
| 17 | 17-2041.00 | Chemical Engineers | 0.36 |
| 18 | 49-9044.00 | Millwrights | 0.36 |
| 19 | 15-2051.02 | Clinical Data Managers | 0.36 |
| 20 | 17-3021.00 | Aerospace Engineering and Operations Technologists and Technicians | 0.35 |

Table 8: Top 20 occupations by (rescaled) similarity (based on LS patents full-texts).