Dipartimento di Economia Internazionale, delle Istituzioni e dello Sviluppo (DISEIS)

WORKING PAPER SERIES

WP 2401

August 2024

Towards the Terminator Economy: Assessing Job Exposure to AI through LLMs

Emilio Colombo, Fabio Mercorio, Mario Mezzanzanica, Antonio Serino

Prior to being published, DISEIS working papers are subject to an anonymous refereeing process by members of the editorial board or by members of the department.

DISEIS Working Papers often represent preliminary work which are circulated for discussion and comment purposes. Citation and use of such a paper should account for its provisional character. A revised version may be available directly from the author.

DISEIS WP 2401

DISEIS
Dipartimento di Economia Internazionale, delle Istituzioni e dello Sviluppo
Università Cattolica del Sacro Cuore
Via Necchi 5
20123
Milano

Towards the Terminator Economy: Assessing Job Exposure to AI through LLMs

Emilio Colombo[*]
Università Cattolica
CRISP
University of Milano-Bicocca

Fabio Mercorio [†]
CRISP and DISMEQ
University of Milano-Bicocca

Mario Mezzanzanica [‡]
CRISP and DISMEQ
University of Milano-Bicocca

Antonio Serino [§]
DISMEQ
University of Milano-Bicocca

August 2024

[*]Corresponding author, Email: emilio.colombo@unicatt.it.
[†]Email: fabio.mercorio@unimib.it
[‡]Email: mario.mezzanzanica@unimib.it
[§]Email: a.serino3@campus.unimib.it

Abstract

There is no doubt that AI and AI-related technologies are reshaping jobs and related tasks, either by automating or by augmenting human skills in the workplace. Many researchers have tried to estimate if, and to what extent, jobs and tasks are exposed to the risk of being automatized by state-of-the-art AI-related technologies. Our work tackles this issue through a data-driven approach: (i) developing a reproducible framework that uses several open-source large language models to assess the current capabilities of AI and robotics in performing work-related tasks; (ii) formalising and computing a measure of AI exposure by occupation, namely the `TEAI` (`T`ask `E`xposure to `AI`) index. Our `TEAI` index is positively correlated with cognitive, problem-solving and management skills, while is negatively correlated with social skills. Our results show that about one-third of U.S. employment is highly exposed to AI, primarily in high-skill jobs, requiring graduate or postgraduate level of education. Using 4-year rolling regressions, we also find that AI exposure is positively associated with both employment and wage growth in the period 2003-2023, suggesting that AI has an overall positive effect on productivity.

Keywords: AI, Large Language models, employment, skills.
JEL codes: J24, O33, O36.

1 Introduction

The famous 1984 movie "The Terminator" is set in a dystopian future where intelligent machines, created by a military defence system known as Skynet, become self-aware and perceive humanity as a threat, initiating a war to eliminate humans. Skynet creates advanced humanoid robots called Terminators to hunt down and kill human survivors. The Terminator has advanced learning algorithms that enable it to adapt to any environment, making it a formidable antagonist for humans. The debate and concerns about the impact of AI are often framed against the backdrop of the film's setting. This paper takes these concerns seriously by developing an AI-centred assessment of the potential exposure of different occupations to artificial intelligence.

Assessing the impact of technology on the labour market is not easy, as there are several potential channels at work. As stressed by Acemoglu and Restrepo (2019), technology has three major effects on labour demand. The first is the productivity effect, which operates through lower production costs brought about by new technologies. The second is the displacement effect of workers by machines and the like. These two effects operate in different directions and depend on whether technology substitutes for or complements human labour; in economic jargon, they depend on the elasticity of substitution between tasks. In addition, there is a third effect of technology: the creation of new tasks and activities in which labour can be productively employed (the reinstatement effect). Indeed, if we look at history, the reinstatement effect has been a central feature of all technological revolutions, which have continually created new opportunities for labour. For the reinstatement effect to take place, technology must have an impact wider than its narrow scope, with spillover effects in sectors/areas other than those for which it was designed. In other words, technology must have the features of a general purpose technology, which, according to Lipsey et al. (2005), are pervasiveness across the economy, ability to generate complementary innovations, and improvement over time.

AI provides strong arguments for being considered a general-purpose technology due to its broad applicability, potential productivity gains and possibility to drive further innovation. However, these characteristics of AI create a relevant measurement problem, as it is extremely difficult to identify all the channels through which it affects the economy. This paper contributes to this field by developing a methodology for assessing AI exposure using Large Language Models (LLMs), with a granular approach that analyses exposure for each task within each occupation.

We provide several innovations with respect to the existing literature.

Methodologically, we design and implement a reproducible framework to assess the extent to which existing AI and robotics technologies can perform job-related tasks based on Large Language Models (LLMs).

Rather than assessing AI exposure through external benchmarks such as expert judgement or AI patent and innovation data, we construct an *internal* assessment using LLM's own evaluation. In other words, LLMs assess their ability to perform each of the 19K job related tasks associated with about 1k occupations as described by O*NET[1]. The main advantage of this approach is that it is fully transparent. Not only does it allow full reproducibility,[2] but more importantly, it allows comparability over time. AI is an innovative and fast-growing field where new models are expected to significantly outperform existing ones. New LLMs can be applied to our approach to measure the extent of their improvement in task performance.

From an economic perspective, we develop an AI exposure measure for all O*NET occupations based on individual tasks .

Our AI exposure index is positively correlated with cognitive, problem-solving and management skills, consistent with the evidence that AI affects management and decision-making tasks; on the other hand, AI exposure is negatively correlated with social skills, where AI has clear limitations.

Bringing our measure to the data, we show that, in the US, approximately 1/3 of employment is highly exposed to AI technologies; most of this employment (88%) is in high skill jobs, requiring graduate or postgraduate level of education. In terms of labour market outcomes, using 4-year rolling regressions, we find that AI exposure is positively associated with both employment and wage growth in the period 2003-2023, suggesting that AI has an overall positive effect on productivity.

The remainder of the paper is structured as follows: section 2 discusses the related literature, section 3 describes the methodology and the construction of the AI index, section 4 presents the results; finally, section 5 concludes.

[1] O*NET: Occupational Information Network, is a comprehensive database containing detailed information on hundreds of standardised and occupation-specific descriptors. It is sponsored by the U.S. Department of Labor/Employment and Training Administration. O*NET serves as a resource to provide information on skills, abilities, knowledge, work activities, and interests associated with occupations

[2] All codes are available on GitHub at https://github.com/Crisp-Unimib/Terminator-Economy

2 Background and Related Works

AI and jobs Since the seminal paper by Autor et al. (2003), the task approach has proved very effective in analysing the impact of technology on jobs. It divides work activities into tasks, each of which can be performed by humans or by machines. In this way, the distinction between capital and labour tasks is more precise, flexible, and able to shift over time. Indeed, capital and machines can substitute for labour in the performance of some tasks, while complementing it in the performance of others.

The task approach has been applied to analyse the effect of technology and trade (offshoring) (Acemoglu and Autor, 2011), to the long run effect of technology (Consoli et al., 2023) and to skill-task interaction (Colombo et al., 2019).

More recently this approach has been used to measure occupational exposure to computers and robots. In a seminal paper Frey and Osborne (2017) estimated that up to 47% of jobs in the US are at risk of automation.[3] Subsequently, other attempts focused on developing measures of exposure to machine learning and robotics (Brynjolfsson and Mitchell, 2017; Acemoglu and Restrepo, 2020) and to AI (Felten et al., 2021; Webb, 2023; Eloundou et al., 2023; Pizzinelli et al., 2023).

Overall, these works find an extensive share of employment exposed to AI; the specific impact on occupations varies depending on the nuances that different indicators of the impact of technology capture, i.e., whether they focus on aspects of technology that impact more routine-based activities (Frey and Osborne, 2017) or more cognitive elements (Felten et al., 2021). A common feature of these papers is the attempt to quantify AI exposure through an *external* benchmark, which may be expert judgement or data analysis on patents and innovations. In contrast, our approach is based on an *internal* assessment, where LLM systems are asked to assess the suitability of tasks for AI. This approach has two main advantages. First, it is fully transparent, with results and outcomes being fully disclosed. Second, the approach is fully reproducible. This means that whenever new generations of LLM are available, they can be used in our approach to measure the change in task exposure that they imply.

Large Language Models LLMs are powerful computational models designed to understand and generate human-like text by harnessing vast amounts of textual

[3] See also Nedelkoska and Quintini (2018) for a similar approach.

data, and have taken Natural Language Processing (NLP) by storm, achieving state-of-the-art performance in many tasks (Min et al., 2023). Typically these models are based on Transformer architecture (Vaswani et al., 2017), powered by Attention mechanism (Luong et al., 2015; Bahdanau et al., 2014) and are composed by decoder-only stack.

These models are initially trained on Autoregressive task (Radford et al., 2018), where given a sequence of words $S = (w_1, w_2, ..., w_{n-1})$ the training objective is to maximize the log-likelihood $\sum_i \log P(w_i|w_1, w_2, \ldots, w_{i-1}; \theta^T)$ where θ^T are the model parameters, in order to predict the next word in the sequence $\prod_{i=1}^{n} P(w_i \mid w_1, ..., w_{i-1})$. After being pre-trained, these models are fine-tuned for several tasks, providing examples of Natural Language Inference (Radford et al., 2018). Thanks to their capability to learn from context, known as in-context learning (Radford et al., 2019), LLMs can accomplish specific tasks with high accuracy (Zhao et al., 2023), exploiting prompt engineering methodologies such as zero-shot (Wei et al., 2021) and few-shot learning (Brown et al., 2020).

3 Building the AI Exposure Index

Our method can be summarised as follows: First, we obtain from O∗NET the description of each task associated with each SOC occupation. Second, we apply LLMs to the task descriptions to obtain a rating of how well AI technologies can perform each task. Third, we aggregate the rating at the occupation level to obtain an AI exposure score. Finally, we apply our score to US data to assess the extent of AI exposure in the US labour market and the impact of AI on employment and wages. Figure 1 provides a graphical representation of our approach.

3.1 Step 1: Compute the AI rate

In order to avoid the risk of being driven by the well-known problem of LLMs called ”hallucinations” (Ji et al., 2023), we design a framework involving three different LLMs aimed at identifying and limiting the false information generated, creating a consensus system between them.

Model choice To ensure the reproducibility of this work we use three of the best open source models, according to performance benchmarks, available on the open

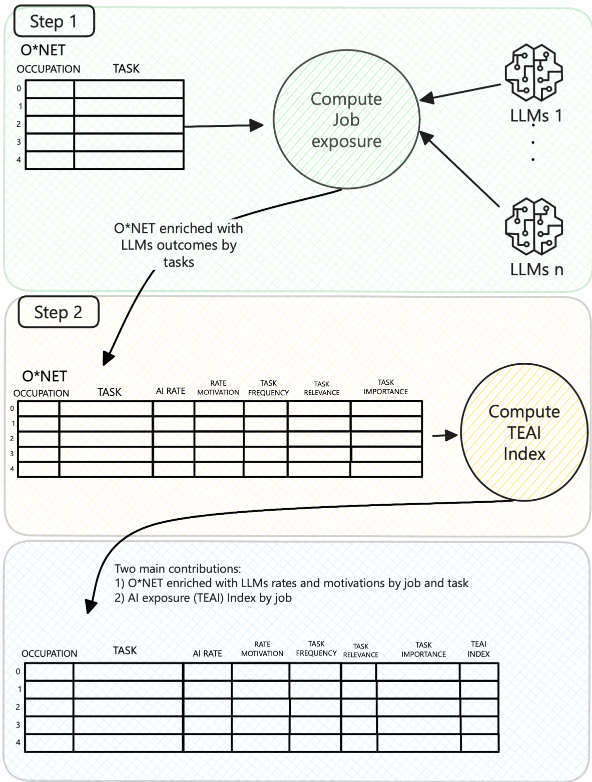


Figure 1: Graphical overview of the framework to compute the `TEAI` Index

LLM leaderboard.[4] To reduce the lack of computational complexity, we use 7 billion parameter models. The three selected models are *Mistral 7B Instruct v 0.2* (Jiang et al., 2023), *openchat 3.5 0106* (Wang et al., 2023) and *orca mini v3 7b* (Mukherjee et al., 2023).

Prompt design The starting point is the O*NET taxonomy which identifies 19281 tasks for 923 SOC occupations.[5] We formulate a five-shot prompt using the few-shot learning approach (Brown et al., 2020). We use each individual task description assigned to an occupation to ask the models how well, on a scale of 1 to 5,[6] the combination of different AI technologies could perform the input task and provide a discursive motivation for the evaluation. As AI technologies, we consider i) LLMs for textual data understanding, ii) Image Processing Systems for elaboration and decision-making based on visual data analysis, and iii) Robotic systems for physical execution. At the end of the prompt, the model is provided with five examples of the task to obtain more contextual and accurate results.

[4] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

[5] We use the O*NET 28.2 version released in February 2024.

[6] How well an AI system, which can be an LLM, Image Processing System or a Robot, could perform in the task on a scale of 1 to 5 where 1 stands for *poor* and 5 stands for *excellent*?

Consensus System We iterate this process for each task provided by O*NET and for each model chosen, ending up with three scores and natural language motivations provided by each model. Table 1 provides an example of the results after this stage for a selection of occupations and tasks.

As mentioned above, the decision to use three different models was made to avoid hallucinations. In order to construct a single indicator, we took a conservative approach by assigning to each task the value of the rating with the highest frequency among the three models; if the three rates were different, we chose the lowest.

To assess the agreement between the three rates expressed by the LLMs, for each single task we compute a consensus metric (Tastle and Wierman, 2007).

$$Cns(a_i) = 1 + \sum_{k=1}^{m} p_k \log_2(1 - \frac{|LV_k - \mu_{LV}|}{d_{LV}}) \tag{1}$$

Equation (1) shows the consensus calculation in which, given an input task a_i, for which three rates are expressed by three different LLMs, LV_k denotes the rate (on scale from 1 to 5) expressed by the k-th LLM, p_k denotes the relative frequency associated with LV_k; μ_{LV} defines the weighted average of the three LV rating values computed using p_k probabilities as weights; finally d_{LV} represents the range of the rating scale adopted.

Specifically, for each k-th rate from 1 to 3 (corresponding to the three available rates), the absolute difference between the rating value LV_k and the weighted average μ_{LV} is determined. This difference is then normalised by dividing by the scale size d_{LV}. The logarithmic function $\log_2$ is applied in order to moderate the impact of larger discrepancies from the mean.

This process allows the three rates to be combined into a single measure of agreement. A higher consensus value indicates greater agreement between the observed rating values expressed for each task by the three different LLMs.

Analogously, to estimate the similarity between the motivations provided by the LLMs, we compute the centroid of semantic cosine similarity (Rahutomo et al., 2012), between them. The embedding vectors for the centroid computation is obtained using an open source Transformer model: as for the LLMs, we chose the Transformer model to be used in accordance with the Massive Text Embedding Benchmark (MTEB) Leaderboard.[7]

[7]https://huggingface.co/spaces/mteb/leaderboard. Having English-language motivations, the choice fell on the *UAE-Large-V1* https://huggingface.co/WhereIsAI/UAE-Large-V1, which rep-

Table 1: Extraction of two occupation profiles and two tasks from O*NET, enriched as discussed in Step1 to derive to what extent existing AI and robotics technologies can perform job-related tasks according to Mistral, Orca-mini and Open-Chat LLMs. Key activities are highlighted for a better reading

O*NET Occupation	**O*NET Task**	**Ratings**			**Motivation provided by Mistral**	**...**
		Mistral	**Orca-mini**	**Open-Chat**		
33cmAdministrative Services Managers	Direct or coordinate the supportive services department of a business, agency, or organization.	4	4	3	Robotics can automate repetitive tasks such as data entry, document scanning, and mail sorting. Image Processing Systems can analyze documents and emails for information extraction and categorization. LLMs can handle complex tasks like scheduling, coordinating resources, and communicating with team members. Combining these AI technologies can significantly improve the efficiency and accuracy of administrative services, allowing managers to focus on more strategic tasks.	...
	. . .					
	Conduct classes to teach procedures to staff.	2	3	2	LLMs could potentially create instructional materials, but they lack the ability to engage and assess staff performance during classes. Image Processing Systems have no role in this task. Robotics could theoretically conduct classes through pre-recorded videos, but the human interaction and customization required for effective training make this a poor solution.	...
33cmTaxi Drivers	Complete accident reports when necessary.	3	2	3	Image Processing Systems can analyze accident scene images to determine the extent of damage and potential causes, while LLMs can assist in generating reports based on the data. However, the complexity of accident reports, which often require a nuanced understanding of legal and insurance terminology, may limit the effectiveness of AI in this task. Robotics could potentially automate the collection of physical evidence, but their role is limited compared to the other AI technologies.	...
	. . .					
	Vacuum and clean interiors and wash and polish exteriors of automobiles.	1	2	1	Robotics are best suited for this task, as they can physically perform the vacuuming, cleaning, and washing functions. LLMs and Image Processing Systems have limited applicability, as they do not possess the physical capabilities required to accomplish the task.	...

The rating scale expressed by the LLMs corresponds to the following categorical labels:
1: Poor, 2: Fair, 3: Average, 4: Good, 5: Excellent

Table 2 shows the results following the calculation of semantic similarity and consensus between the different rates. In particular, since higher cosine similarity values reflect higher semantic similarities between the text of the LLM motivations, we expect a strong correlation between the consensus metric and cosine similarity. The heat map represented in figure 2 shows that both the values of cosine similarity and the consensus metric are extremely high, with an average close to 0.9 in both cases.

Table 2: The two occupation profiles and two tasks from O*NET, shown in Table 1, enriched with the centroid of semantic similarity and consensus measure.

O*NET Occupation	**O*NET Task**	**...**	**Similarity**	**Consensus**
Administrative 2*Services Managers	Direct or coordinate the supportive services department of a business, agency, or organization.	...	0.918	0.828
	⋮			
	Conduct classes to teach procedures to staff.	...	0.935	0.828
2*Taxi Drivers	Complete accident reports when necessary.	...	0.948	0.828
	⋮			
	Vacuum and clean interiors and wash and polish exteriors of automobiles.	...	0.925	0.828

On the one hand, this suggests coherence between LLM-generated rates and the associated motivations, and on the other hand it adds robustness to our conservative approach in choosing between different models.

The result of this process is a single score TE that returns a metric from 1 to 5 measuring the extent to which AI can perform each specific task, and a quantitative indicator of the similarity between the discursive motivations generated by the models.

3.2 Step 2: Compute the AI exposure

To compute occupation exposure to AI, we follow Felten et al. (2021) and aggregate the TE scores at the occupation level by weighting them by task relevance (R), im-

resented an excellent compromise between effectiveness and efficiency, given its small size.

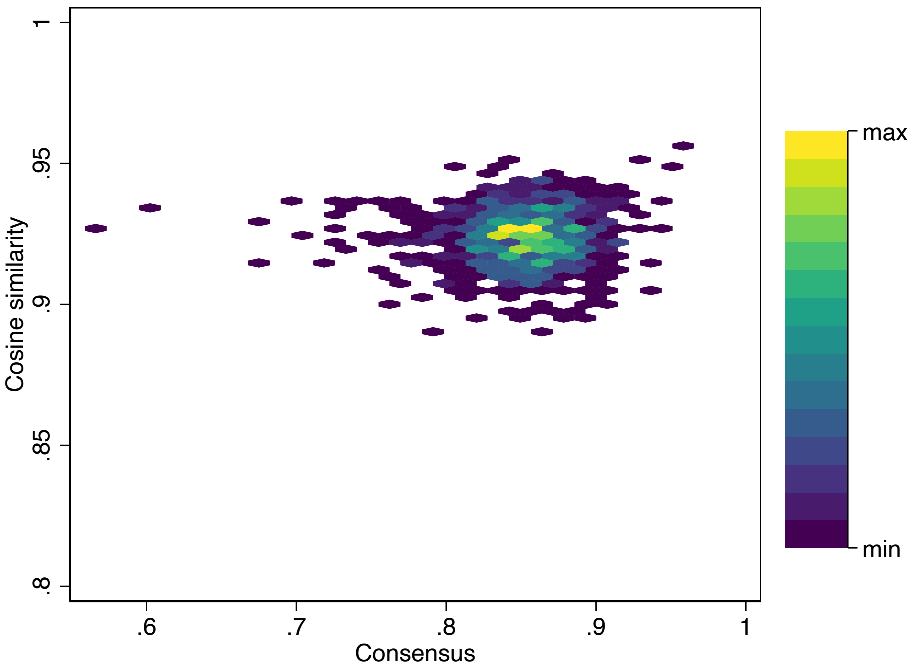


Figure 2: Heat map between cosine similarity of textual motivation of LLMs and consensus measure between scores. Data are aggregated at occupation level

portance (I), and frequency (F) as measured by O*NET.[8] More specifically, for each task j and occupation i our AI exposure score is computed as follows:

$$TEAI_i = \frac{\sum_{j=1}^{n} TE_{ij} \cdot R_{ij} \cdot I_{ij} \cdot F_{ij}}{\sum_{j=1}^{n} \cdot R_{ij} \cdot I_{ij} \cdot F_{ij}} \tag{2}$$

where TE_{ij} identifies the metric developed in step 1 at task level, n defines the number of tasks within each occupation. Each weight is scaled by its maximum to obtain equal weights. The O*NET model uses different scales for Relevance (scale 1-100), Importance (scale 1-5), and Frequency (scale 1-7). We normalise the indexes to ensure equal scale across weights. Finally, the score is normalised to ensure comparability with other similar scores.

[8]These weights capture different aspects of the tasks. More specifically. Importance: indicates the degree to which a particular descriptor is important for the occupation. Relevance refers to the proportion of incumbents who rated the task as relevant to their job. Frequency refers to the frequency of each task within the occupation, from annual to hourly. Despite providing task descriptions, O*NET does not provide rating, importance and frequency of tasks for 39 occupations. We manually assigned values for them.

4 Experimental Results

4.1 Benchmarking evaluation

First, we compare our AI index with other existing measures in the literature. Figure 3 shows the correlation between the TEAI index and the well-known measure developed by Frey and Osborne (2017), the AI exposure index by Felten et al. (2021) and by Webb (2023), and the offshorability index developed by Acemoglu and Autor (2011). The pairwise correlation is always statistically significant at 5%. It is higher for the AIOE index, much lower for the AI Webb and the offshorability index, and negative for the Frey-Osborne index. This means our measure is broadly consistent with existing measures but captures different elements of the relationship between AI and the labour market. The negative correlation with the Frey and Osborne index can be explained by the fact that the latter is a measure of exposure to robotics and computerisation and is more focused on routine tasks. At the same time, generative AI is more focused on cognitive/non-routine tasks.

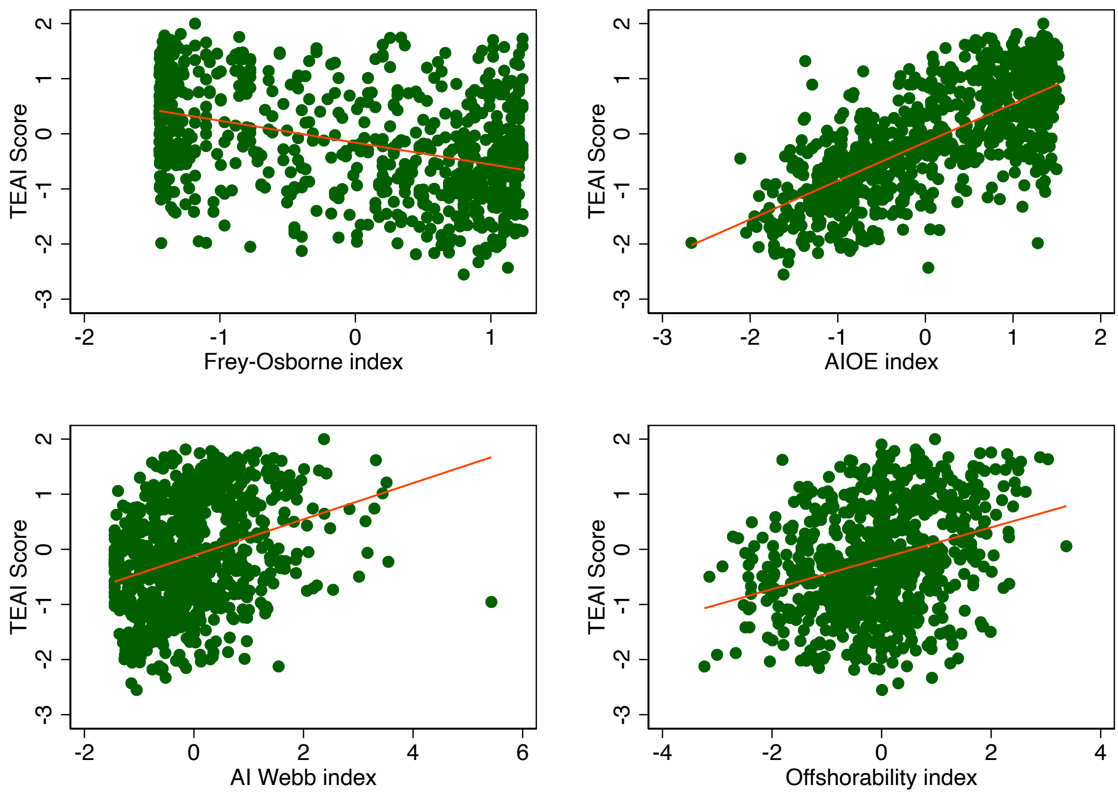


Figure 3: Correlation with existing exposure indexes. Each dot represents a SOC occupation.

4.2 AI and skills

Next, we explore the relationship between our `TEAI` index and different skills. Figure 4 displays scatterplots comparing the `TEAI` index with the intensity of different skill types at occupation level derived from Acemoglu and Autor (2011). The graph shows the peculiar nature of AI technologies, which are positively correlated with cognitive analytical and interpersonal skills, but negatively correlated with routine manual skills and non-routine manual skills that require physical adaptability. Surprisingly, the correlation with cognitive routine skills is only weakly positive, while it is positive for non-routine manual skills requiring interpersonal adaptability.

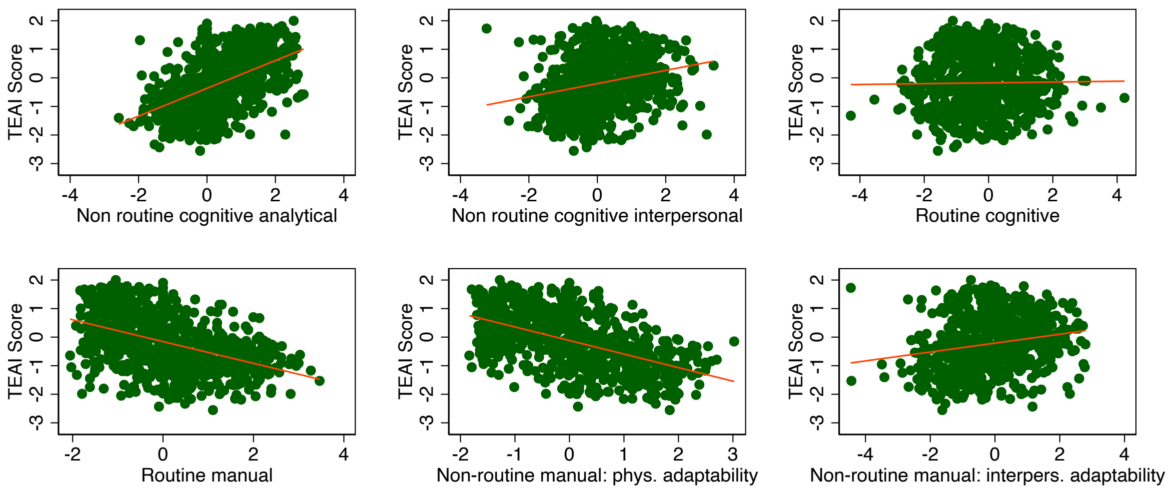


Figure 4: Correlation with different skill intensity measures. Each dot represents a SOC occupation.

The results of the figure are purely descriptive, therefore we add a more robust analysis by extracting from O*NET the detailed skills associated with each occupation. We group the skills into 4 classes: Cognitive, Social, Problem Solving and Management, and Technical skills. We then develop a skill relevance index for each class at the occupation level by weighting each skill according to its level and importance.[9] The skill relevance index is constructed as follows:

$$SR_{ci} = \frac{\sum_{z=1}^{m} S_{zcj} \cdot L_{zcj} \cdot I_{zcj}}{\sum_{z=1}^{m} \cdot L_{zcj} \cdot I_{zcj}} \tag{3}$$

where z denotes the m skills of class c in each occupation j; L and I denote, respectively, the level and importance of each skill in each occupation.

[9]As provided by O*NET.

We then estimate the following regression:

$$TEAI_i = \alpha + S_i'\beta + \gamma O_i + \epsilon_i$$

where each observation is a SOC occupation (i), $TEAI_i$ is our measure of AI exposure, S is a 4x1 column vector of skill relevance for each of the skill classes described above at the occupation level, β is a 4x1 column vector of coefficients, O defines occupation dummies and ϵ_i the error term. We saturate the model with more detailed dummies up to the fifth digit; therefore, the results are identified within each group variation. Table 3 shows the results. The `TEAI` index is positively correlated with cognitive, problem solving and managerial skills, but negatively correlated with social skills, as expected. The relationship with technical skills is very weak and does not survive the inclusion of detailed SOC occupation dummies.

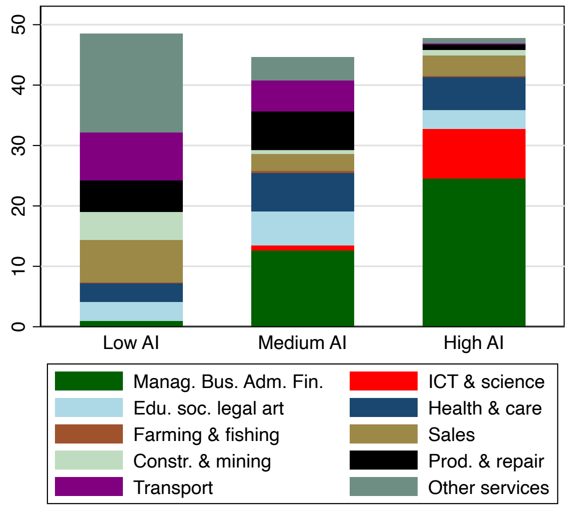
Table 3: OLS estimates of TE-AI index on measures of skill intensity

	(1))	(2)	(3)
Cognitive	5.6934***	5.3653***	5.3105***
	(0.8023)	(0.7937)	(1.1712)
Social	-3.1395***	-3.2464***	-3.0310**
	(0.7239)	(0.7335)	(1.1295)
Prob. sol. man.	3.1883***	3.4810***	2.9411*
	(0.8183)	(0.8129)	(1.3644)
Technical	0.0259	-0.0419	-0.1486
	(0.6432)	(0.6527)	(0.8556)
SOC FE	3d	4d	5d
R2	0.775	0.780	0.854
N	774	771	522

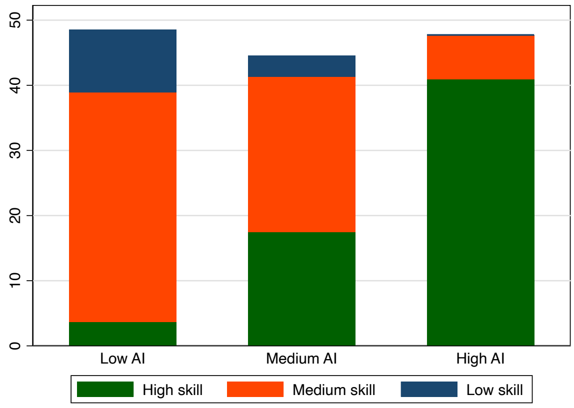
Source: Authors' calculation on O*NET and BLS data.
Note: Each observation consists of an occupation. OLS regression using `TEAI` index as the dependent variable. The independent variables are skill intensities. All the regressions include occupation (SOC) fixed effects at 3, 4 and 5 digits. Robust standard errors in parentheses *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

4.3 AI employment and wages

Finally, we explore the relationship between `TEAI` and labour market outcomes. We start by analysing the size and the characteristics of workers exposed to AI technologies. First, we divide the distribution of TEAI scores into three tertiles representing high, medium and low AI exposure. We then compute the degree of exposure of the US population using BLS employment data. Finally, we distinguish between occupational groups and by skill groups within each tertile.



(a) TEAI index by SOC group



(b) TEAI index by skill intensity

Figure 5: Exposure to TEAI index by SOC group (Fig. 5a) and by skill intensity (Fig. 5b). US BLS employment. Values in millions of workers. Each bar represents a tertile of the TEAI score distribution.

Figures 5a and 5b show the results. Overall, in 2023, 34% of US employment is highly exposed to AI technologies, while medium and low exposure represents 32% and 34%, respectively. Our findings do not suggest a polarising effect of AI exposure as found by Frey and Osborne (2017); on the contrary, AI seems to have a more balanced impact on the labour market. This is because our indicator is able to capture recent advances in AI, such as LLMs, which have affected occupational groups such as management, business, administration and finance, as well as ICT and science, which are intensive in non-routine cognitive tasks. For example, AI technologies are increasingly being used to diagnose diseases, write reports, code or brainstorm ideas in management and business. On the contrary, previous studies that focus more on the impact of AI on routine tasks find that these tasks and occupations are less exposed to AI.

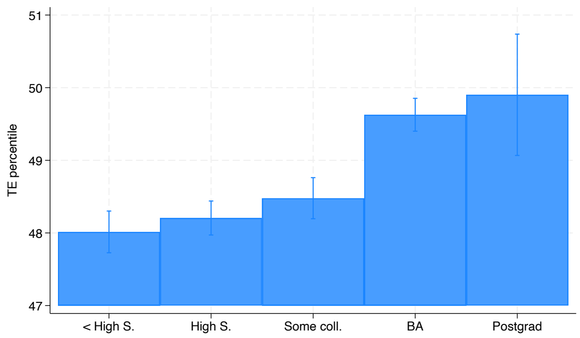
Grouping occupations by skill intensity shows that in the group with high exposure to AI, 88% of employment is in high-skill jobs; in the group with medium exposure 53% of employment is in medium-skill jobs while 40% in high-skill jobs. In the group with the lowest exposure to AI, 67% are medium-skill jobs and 25%

low-skill jobs. Overall, AI exposure disproportionately affects high-skill jobs, which are characterised by the competencies most heavily affected by AI technologies.

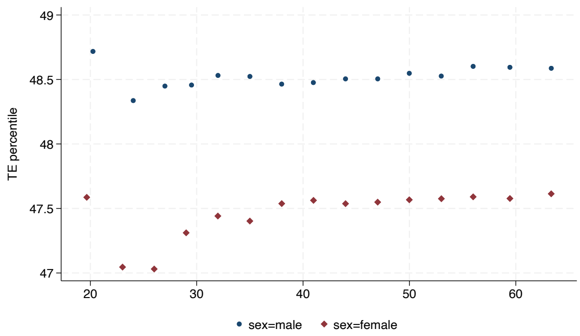
Next we analyse the relationship between AI exposure and workers' characteristics.

Figure 6 shows that TEAI exposure is higher for workers' with a high level of education, particularly graduates and postgraduates. There is a slight increase in exposure with age, although the variation in exposure is really limited after the age of 30. Men are more exposed than women at all ages.

Figure 6: AI exposure by workers' characteristics.



(a) Exposure by education



(b) Exposure by age and sex

Panel a) shows coefficients of regression of education categories on TEAI exposure (in percentiles). Covariates include age and sex. Estimates control for occupation(4d), industry(3d), state and year fixed effects. ACS individual weights are used. Robust standard errors are clustered at the industry level. Panel b) is a binscatter. The x-axis is the average age of workers in an industry-occupation-state observation in the 2022-18 ACS 5 years sample. Biscatter is computed considering education as a covariate. ACS individual weights are used.

Finally, we assess the relationship between AI exposure, employment, and wages. To compute the medium-term effect of AI in a flexible way, allowing for changes during the estimation period, we compute the log change in employment and wages over a 4-year rolling window from 2003 to 2023.[10] Therefore, for each 4Y window, we run the following regression:

$$\Delta y_{i,j} = \alpha + \beta TEAI_i + Z'_{i,j}\gamma + \delta_i + \eta_j + \epsilon_{i,j} \tag{4}$$

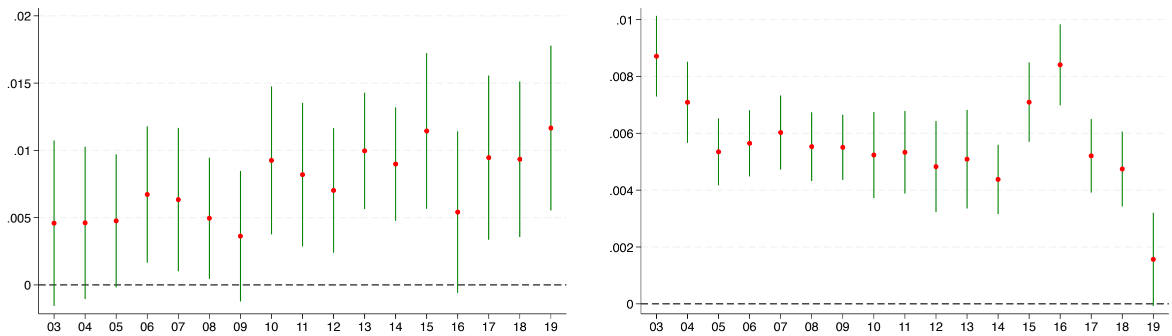
where $\Delta y_{i,j}$ denotes the 4-year change in log employment and log wages in sector j for occupation i, Z is a column vector of controls, δ_i denote occupation dummies, η_j sector dummies and $\epsilon_{i,j}$ the error term. To control for possible endogeneity and

[10]The last 4Y variation is therefore 2019-2023.

omitted variable problems, we include as controls the initial level of employment, the initial level of wage, and wage squared. We also include detailed sector (NAICS) and occupation (SOC) fixed effects in the regression and cluster the errors at the sector level. Figure 7 shows that exposure to artificial intelligence (AI) positively correlates with employment and wage growth throughout the entire period. This suggests that AI technologies complement labour and increase productivity, thereby boosting employment and wages in occupations with greater exposure to AI.

The presence of detailed controls at the industry and occupation level allows us to control for factors on the production side (changes in output across industries), on the demand side (changes in product demand across industries) and on the labour supply side (changes in employment across industries and occupations) that are unrelated to AI technologies and that could affect wages and employment. Moreover, the focus on a relatively short period of time isolates our results from long-term trends within industries and occupations.

Figure 7: `TEAI` index, employment and wage growth.



(a) `TEAI` index and employment growth

(b) `TEAI` and wage growth

This figure plots the effect of AI score on employment and wage growth. Estimates are obtained from equation 4, with rolling regression coefficients and 95% confidence intervals of 4-year windows, starting in 2003-2007 and ending in 2019-2023. The point estimate refers to `TEAI` score, and the dependent variables are annual percentage growth rates of employment and wages. Employment regression includes the log of the initial period of employment. Wage regression includes a log of initial period employment, log initial period wage and log initial period wage squared. All the regressions include occupation (SOC 4 digit) and sector (NAICS 3 digit) fixed effects. Robust standard errors clustered at NAICS level.

Therefore, the positive relationship between employment and wages and AI exposure should be interpreted as meaning that occupations more exposed to AI have stronger employment and wage growth within the occupation and sector. Our results contrast with those obtained by Acemoglu et al. (2022); Webb (2023), who find a negative relationship between employment and wages. The potential reconciliation between our findings and theirs lies, on the one hand, in our construction of a different measure of AI exposure that emphasises more recent advances in AI. On the other hand, our analysis focuses on changes over the last 20 years, while their analysis takes a longer-term perspective, focusing on changes over several decades.

5 Conclusions

This paper provides a comprehensive assessment of AI exposure for 19281 tasks for 923 SOC occupations identified by O*NET. We use the task description and perform the task assessment using LLM's own evaluation. We then aggregate the task scores to obtain an occupation-based score of AI exposure. Our methodology ensures full reproducibility of results, allowing future assessment of potential performance improvements in new versions of LLMs. Our index of AI exposure is positively correlated with cognitive, problem-solving and management skills, highlighting the role of recent advances in AI that have a strong impact on management and decision-making tasks; conversely, our measure is negatively correlated with social skills, a known weakness of AI.

Regarding labour market outcomes, we find that AI exposure is positively correlated with both employment and wage growth over the period 2003-2023, suggesting that AI has a positive effect on productivity. Therefore, at least in the medium term, AI has an overall positive impact on the labour market. However, our estimates show that about one-third of the American workforce is at high risk of AI exposure, most of them in high-skilled jobs. For these workers, whether AI will be an opportunity or a threat in the future will depend on whether it complements or substitutes for human labour. Our measure is relatively agnostic in this respect, as we cannot yet disentangle substitutability from complementarity. In other words, a high exposure measure does not necessarily imply full substitution of labour by technology, which on the contrary may fully complement human activities, leading to higher productivity without displacing labour. Future research will explore this important distinction.

References

Acemoglu, D., Autor, D., 2011. Skills, tasks and technologies: Implications for employment and earnings, Elsevier. volume 4 of *Handbook of Labor Economics*, pp. 1043–1171.

Acemoglu, D., Autor, D., Hazell, J., Restrepo, P., 2022. Artificial Intelligence and Jobs: Evidence from Online Vacancies. Journal of Labor Economics 40, 293–340.

Acemoglu, D., Restrepo, P., 2019. Automation and new tasks: How technology displaces and reinstates labor. Journal of Economic Perspectives 33, 3–30.

Acemoglu, D., Restrepo, P., 2020. Robots and jobs: Evidence from us labor markets. Journal of Political Economy 128, 2188–2244.

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

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 .

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

Brynjolfsson, E., Mitchell, T., 2017. What can machine learning do? workforce implications. Science 358, 1530–1534.

Colombo, E., Mercorio, F., Mezzanzanica, M., 2019. AI meets labor market: Exploring the link between automation and skills. Information Economics and Policy 47, 27–37.

Consoli, D., Marin, G., Rentocchini, F., Vona, F., 2023. Routinization, within-occupation task changes and long-run employment dynamics. Research Policy 52, 104658.

Eloundou, T., Manning, S., Mishkin, P., Rock, D., 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. Papers 2303.10130. arXiv.org.

Felten, E., Raj, M., Seamans, R., 2021. Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. Strategic Management Journal 42, 2195–2217.

Frey, C.B., Osborne, M.A., 2017. The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change 114, 254–280.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P., 2023. Survey of hallucination in natural language generation. ACM Computing Surveys 55, 1–38.

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E., 2023. Mistral 7b. arXiv:2310.06825.

Lipsey, R., Carlaw, K.I., Bekar, C.T., 2005. Economic Transformations: General Purpose Technologies and Long-Term Economic Growth. Oxford University Press.

Luong, M.T., Pham, H., Manning, C.D., 2015. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 .

Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, O., Agirre, E., Heintz, I., Roth, D., 2023. Recent advances in natural language processing via large pre-trained language models: A survey. ACM Computing Surveys 56, 1–40.

Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., Awadallah, A., 2023. Orca: Progressive learning from complex explanation traces of gpt-4. arXiv:2306.02707.

Nedelkoska, L., Quintini, G., 2018. Automation, skills use and training. OECD Social, Employment and Migration Working Papers 202. OECD Publishing.

Pizzinelli, C., Panton, A.J., Tavares, M.M.M., Cazzaniga, M., Li, L., 2023. Labor Market Exposure to AI: Cross-country Differences and Distributional Implications. IMF Working Papers 2023/216. International Monetary Fund.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., 2018. Improving language understanding by generative pre-training .

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI blog 1, 9.

Rahutomo, F., Kitasuka, T., Aritsugi, M., et al., 2012. Semantic cosine similarity, in: The 7th international student conference on advanced science and technology ICAST, University of Seoul South Korea. p. 1.

Tastle, W.J., Wierman, M.J., 2007. Consensus and dissention: A measure of ordinal dispersion. International Journal of Approximate Reasoning 45, 531–545. North American Fuzzy Information Processing Society Annual Conference NAFIPS '2005.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Wang, G., Cheng, S., Zhan, X., Li, X., Song, S., Liu, Y., 2023. Openchat: Advancing open-source language models with mixed-quality data. arXiv preprint arXiv:2309.11235 .

Webb, M., 2023. The impact of artificial intelligence on the labor market. Mimeo, Stanford University.

Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V., 2021. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 .

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al., 2023. A survey of large language models. arXiv preprint arXiv:2303.18223 .