EasyChair Preprint

№ 2451

Does AI Qualify for the Job? A Bidirectional Model Mapping Labour and AI Intensities

Fernando Martínez-Plumed, Songül Tolan, Annarosa Pesole, Jose Hernandez-Orallo, Enrique Fernández-Macías and Emilia Gómez

February 4, 2020

Does AI Qualify for the Job?
A Bidirectional Model Mapping Labour and AI Intensities

Fernando Martínez-Plumed
Universitat Politècnica de Valencia &
European Commission JRC
fmartinez@dsic.upv.es

Songül Tolan
European Commission JRC
songul.tolan@ec.europa.eu

Annarosa Pesole
European Commission JRC
annarosa.pesole@ec.europa.eu

Jose Hernández-Orallo
Universitat Politècnica de Valencia
jorallo@upv.es

Enrique Fernández-Macías
European Commission JRC
enrique.fernandez-macias@ec.europa.eu

Emilia Gómez
European Commission JRC
emilia.gomez-gutierrez@ec.europa.eu

Abstract

In this paper we present a setting for examining the relation between the distribution of research intensity in AI research and the relevance for a range of work tasks (and occupations) in current and simulated scenarios. We perform a mapping between labour and AI using a set of cognitive abilities as an intermediate layer. This setting favours a two-way interpretation to analyse (1) what impact current or simulated AI research activity has or would have on labour-related tasks and occupations, and (2) what areas of AI research activity would be responsible for a desired or undesired effect on specific labour tasks and occupations. Concretely, in our analysis we map 59 generic labour-related tasks from several worker surveys and databases to 14 cognitive abilities from the cognitive science literature, and these to a comprehensive list of 328 AI benchmarks used to evaluate progress in AI techniques. We provide this model and its implementation as a tool for simulations. We also show the effectiveness of our setting with some illustrative examples.

Introduction

In this paper we present a setting for the analysis and simulation of the *intensity* flows between Artificial Intelligence (AI) research and the labour market. Intensity is understood as the relevance of and effort spent on any undertaking. For instance, in the case of an occupation one can estimate how much time a particular activity requires. In the case of AI, one can estimate how much effort (in terms of activity) is devoted to a certain task in a particular area of research. Without a model, some direct connections can be made, such as the observation that progress in machine translation will have an impact on human translators, or that in order to rationalise the cost in language translation and subtitling of a major video-on-demand company, more progress of AI in this area would be needed. But the connections become more complex when we wonder how much AI research in natural language processing is affecting a lawyer, or what areas in AI should require more activity to alleviate the bottleneck of auditors, or any other profession. A traceable two-way analysis would be a more anticipatory and prescriptive analysis than just predicting what jobs are more suitable of automation, assuming things equal or extrapolating from a predictive model in which we cannot have any intervention. A model mapping labour and AI research that allows for counterfactuals could account for the relation between AI and labour in ways that could better represent different scenarios and guide policies according to them.

Differently from previous approaches that have tried to link directly AI developments with labour-related task characteristics (Brynjolfsson, Mitchell, and Rock 2018), our framework adds an intermediate dimension of cognitive abilities which gives us greater flexibility as well as a broader understanding on the impact of AI on labour tasks. More precisely, on one side, we map 14 generic cognitive abilities taken from the cognitive science literature to 59 generic labour-related tasks from task-based surveys from the workplace. On the other side, we map these 14 generic abilities to a comprehensive list of 328 benchmarks used to promote and measure the progress in different areas of AI.

In this regard, we start with the detailed set of labour-related tasks (and occupations) from (Fernández-Macías et al. 2016; Fernández-Macías and Bisello 2017; Fernández-Macías et al. 2018), which are assessed according to the cognitive abilities they typically require. Here we link these cognitive abilities to AI intensity indicators in terms of research activity and interest using AI benchmarks (see Figure 1). We also perform a cluster analysis to see how the AI benchmarks group together given the underlying structure of their required cognitive abilities in order to further increase the interpretation of the results.

This mapping between tasks and AI benchmarks allows us to accurately assess how the intensity of AI research may affect work-related tasks and corresponding occupations, as well as the other way around: how task and occupation intensity should be translated to AI research. We then use this setting to rank tasks by potential AI impact, and to show which areas of AI research should be intensified to have an impact in particular selected tasks and occupations. The main contributions of this paper are summarised as follows:

- We propose a formal matrix-based bidirectional setting for the analysis of the impact between AI research and the labour market.
- We show how identifying the specific cognitive abilities that can be performed by AI gives a broader understanding on the impact of AI on labour tasks, and vice versa.

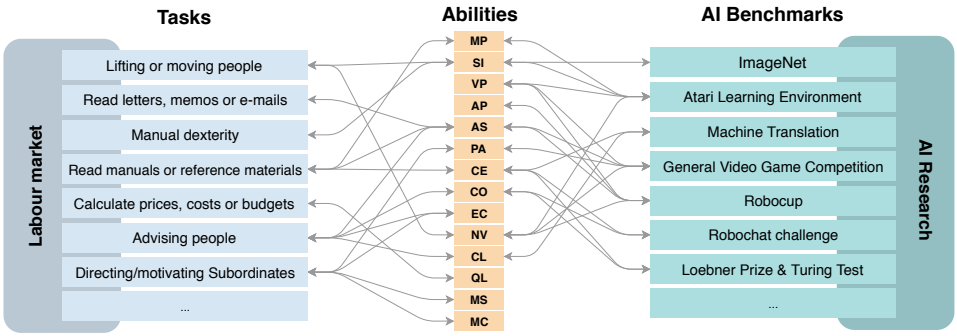

Figure 1: Bidirectional and indirect mapping between job market and Artificial Intelligence. The notation we will use will be $\mathbf{t}$ for the tasks, $\mathbf{a}$ for the abilities and $\mathbf{b}$ for the benchmarks. The arrows are represented by correspondence matrices $\mathbf{W}$ (task-ability correspondence) and $\mathbf{R}$ (ability-benchmark correspondence).

- We see the lack of alignment between the intensities coming from the activity in the workplace and the intensities coming from the activity in AI benchmarks.
- We provide a grouped interpretation of the activity in AI research by performing a cluster analysis on AI benchmarks given the underlying structure of their required cognitive abilities.
- We show how our setting allows for the analysis of counterfactual simulated scenarios and the identification of situations where AI research does not match the required abilities in the labour market.
- We develop an online visual approach[1] for showing the intensity flows between AI benchmarks and the labour market tasks and occupations.

Related Work

The presented setting builds on the labour economics literature focused on measuring the potential for automation on the labour market (Frey and Osborne 2017; Arntz, Gregory, and Zierahn 2016; Nedelkoska and Quintini 2018; Brynjolfsson, Mitchell, and Rock 2018). However, we have to draw a clear line between the impact and technological feasibility of AI to modify the workplace and the configuration of tasks and occupations, and a more simplistic view of AI as leading to full automation (substitution through machines) (anonymous). With this paper, we further complement the literature with a formal setting for measuring AI potential in cognitive abilities and, subsequently, in labour-related tasks and occupations. On the AI side, we perform this by relying on AI benchmarks, as used by researchers and industry to encourage and evaluate progress in AI, instead of relying on expert predictions on the future automatability of occupations, as in (Frey and Osborne 2017) and subsequent studies. This is also in contrast to the use of models that quantify the probability of computerisation for different occupations based on their proportion of routine and non-routine tasks (Autor, Levy, and Murnane 2003). Furthermore, we complement Brynjolfsson et al.'s measure of "suitability for machine learning" for labour-related tasks (Brynjolfsson, Mitchell, and Rock 2018), which draws upon particular technologies in machine learning only. Here, we use a more comprehensive list of AI tasks and benchmarks (which can be further extended and updated to future developments).

The use of AI benchmarks to analyse the state of the art of AI research has been popularised by the seminal work done by the Electronic Frontier Foundation (EFF) (Eckersley, Nasser, and others 2017), and reports such as the AI Index (Shoham et al. 2018), which also covers jobs briefly. Using the EFF data, (Felten, Raj, and Seamans 2018) make a more explicit connection with the labour market. They measure progress in AI through linear trends in benchmarks across different metrics. However, due to nonlinear performance jumps at certain thresholds of each benchmark, progress in different benchmarks cannot be measured in a comparable manner. We address this issue by translating benchmarks to a measure of AI research activity, and not the incommensurate magnitudes of each benchmark. (Felten, Raj, and Seamans 2018) introduce abilities, but they are specialised for "job task requirements", which limits its independence to the labour connection, and precludes a balanced bidirectional analysis.

In this paper, we integrate several theories of intelligence and cognition in psychology, animal cognition and AI textbooks to give a broader definition of abilities, as a more independent latent layer than human abilities (work-oriented) or AI abilities (technology-oriented). We draw information from a very comprehensive set of AI benchmarks, competitions and tasks (see section for details), ensuring a broad coverage of AI tasks. Unlike many of the previous approaches, we formalise our setting by proposing a unified matrix-based mathematical model for the specification of dynamic intensities for AI and labour tasks. This formalisation allows for the analysis of intensity flows between AI and labour tasks (in both directions) analytically. This makes it possible to study real scenarios as well as simulated ones, using counterfactual or speculative hypotheses varying the intensity levels across tasks or AI benchmarks.

[1] https://safe-tools.dsic.upv.es/shiny/OTAAI/

Data

For the two extremes of our mapping, as shown in Figure 1, we need to rely on very different sources of data. We start with a description of labour-related task intensity before moving to a description of research intensity in AI.

Tasks and occupations

We gather the data about labour-related tasks and occupations from (Fernández-Macías et al. 2016; Fernández-Macías and Bisello 2017; Fernández-Macías et al. 2018), comprising a list of tasks and their respective intensity (i.e. relevance and time spent) across occupations.

Concretely, we classify occupations according to the 3-digit International Standard Classification of Occupations (ISCO-3)[2]. Since there is no international data source that covers the full classification, we combine data from three different sources: (1) the European Working Conditions Survey (EWCS)[3]; (2) the OECD Survey of Adult Skills (PIAAC)[4]; and (3) the database from the Occupational Information Network (O*NET)[5]. While (1) and (2) are surveys that provide data measured at the individual worker level based on replies to questions on what they do at work, (3) is also based on employer job postings, expert research and other sources. O*NET is widely used in the literature on labour markets and technological change (Acemoglu and Autor 2011; Frey and Osborne 2017; Goos, Manning, and Salomons 2009) and it covers a large share of the task list that we use in our analysis. However, the occupational level of the data precludes a further analysis of the variation in task content within occupations. Moreover, much like the EWCS for Europe, the O*NET is based on US data only. Therefore, even if there are likely differences in the task content of occupations across countries (due to institutional as well as socio-economic differences) we cannot analyse these differences in the present analysis.

In these sources, task intensity for different occupations is derived either as a measure of time spent on specific tasks (e.g., the intensity for the task *"Lifting or moving people"* is obtained from survey question *"Does your main paid job involve lifting or moving people?"* and the corresponding 7-point scale answers ranging from *"All of the time"* to *"Never"*), or curated by occupational experts and provided on a standardised occupational level (e.g., the extent to which the task is required to perform a job). Due to the varying nature of survey data, we need to be aware of issues such as measurement error, high variation in responses across individuals and biased responses. Consistency in the measurement of task intensity across the different data sources is measured with Cronbach's alpha, which is calculated from the pairwise correlation between items that measure similar concepts. All tests yield high correlations and Cronbachs's Alpha values of between 0.8 and 0.9.

Finally, in order to make the measures of task intensity comparable across all three data sources, we equalise scales and levels of all variables. For this purpose, we rescale the variables to a $[0, 1]$ scale with 0 representing the lowest possible intensity and 1 representing the highest possible intensity of each variable. Moreover, we average scores measured on an individual level (i.e., all variables from PIAAC and EWCS) to the unified level of standardised 3-digit occupation classifications. The final database contains the intensity of 59 tasks across 119 different occupations.

AI benchmarks

We consider a comprehensive set of AI benchmarks for our setting based on previous analysis and annotation of AI papers (Hernández-Orallo 2017a; Martínez-Plumed et al. 2018; Martinez-Plumed and Hernandez-Orallo 2018) as well as open resources such as *Papers With Code*[6] (the largest, up-to-date, free and open repository of machine learning papers). It includes data from multiple (verified) sources, including academic literature, review articles and code platforms focused on machine learning and AI.

From the aforementioned sources we track the reported evaluation results on different metrics of AI performance across separate AI benchmarks (e.g., tasks, datasets, competitions, awards, etc.) from a number of AI domains. We cover computer vision, speech recognition, music analysis, machine translation, text summarisation, information retrieval, robotic navigation and interaction, automated vehicles, game playing, prediction, estimation, planning, automated deduction, among others. This ensures a broad coverage of AI tasks, well beyond perception, such as the ability to plan and perform actions on such plans. Specifically, our framework uses data from 328 different AI benchmarks, after selecting those with enough information available for different evaluation metrics.

When aiming at evaluating the progress in specific AI areas, we need to pay attention to the set of criteria about how a system is to be evaluated. Even if the metrics that are used in each benchmark improve, it would be misleading to consider that the progress in AI should be analysed by aggregating these values. First, these magnitudes are incommensurate, so aggregating the score in a video game with the result of translation task is meaningless. Second, the results are obtained by specific systems solving particular tasks. There is no understanding on how to build systems that can solve all these tasks at the same time.

Therefore, instead of using the rate of progress with particular performance metrics, we will analyse the activity level or *intensity* for a benchmark, measured in terms of the production (e.g., outputs such as research publications, news, blog-entries, etc) from the AI community. Benchmarks that have increasing trends in their production rates –not their performance metrics– indicate that more AI researchers and practitioners are working on them (i.e., there is a clear research effort and intensity). Note that this is not an indication of progress, although, presumably, effort may lead to some progress eventually.

[2]https://www.ilo.org/public/english/bureau/stat/isco/

[3]https://www.eurofound.europa.eu/surveys/european-working-conditions-surveys

[4]https://www.oecd.org/skills/piaac/

[5]https://www.onetonline.org/

[6]https://paperswithcode.com/

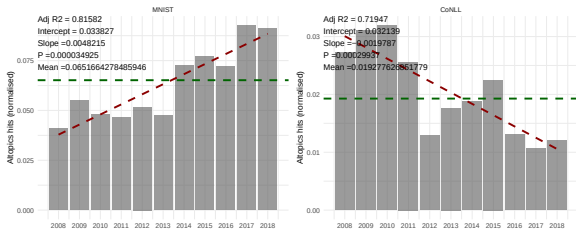

Figure 2: Average rate of activity level or intensity (green dashed line) for a couple of illustrative AI benchmarks over the last decade (2008-2018).

In order to derive the activity level or *intensity*, we will use some proxies. In particular, we performed a quantitative analysis using data obtained from *AI topics*[7], an archive kept by the Association for the Advancement of Artificial Intelligence (AAAI)[8]. This platform contains a myriad of AI-related documents (e.g. news, blog entries, conferences, journals and other repositories from 1905 to 2019) that are collected automatically with NewsFinder (Buchanan, Eckroth, and Smith 2013). In this regard, in order to calculate the intensity in each particular benchmark, we average the number of hits (e.g., documents) obtained from *AI topics* per benchmark and year over a specific period of time. Note that the number of hits are normalised to sum up to 100% per year. Figure 2 shows the activity trends for two different benchmarks. Our measure of intensity is the average over the period 2008-2018.

Model

In the following subsections we describe the main components of our model, as originally illustrated in Figure 1. We will use the following notation:

- $\mathbf{t}$: (labour) task intensity vector.
- $\mathbf{W}$: task-ability correspondence matrix.
- $\mathbf{a}$: ability vector.
- $\mathbf{R}$: ability-benchmark correspondence matrix.
- $\mathbf{b}$: benchmark intensity vector.

We define them in more detail below.

Intensity vectors

Vector $\mathbf{t}$ denotes task intensities. In section "Tasks and occupations" we described the data we will use, meaning that $\mathbf{t}$ will have dimension (59×1), on a $[0, 1]$ scale with 0 and 1 representing the lowest and highest possible intensity respectively. This vector reflects the occupational task intensity in the abilities assigned to the tasks in each occupation (note that each occupation has a different $\mathbf{t}$ vector).

On the other hand, $\mathbf{b}$ denotes a benchmark intensity vector (328×1), with relative values in $[0, 1]$. This vector shows the average (normalised) number of documents obtained from *AI topics* per benchmark and year over a specific period of time as explained in section "AI benchmarks".

[7]https://aitopics.org

[8]https://www.aaai.org/

Cognitive abilities

In previous works (Autor 2013; Acemoglu et al. 2014), labour-related tasks and those that are usually set in AI as capacities are usually matched directly, even if the elements on the left list in Figure 1 are very different from the elements on the right. However, tasks and benchmarks can be mapped through an intermediate layer of latent factors, what we refer to as 'cognitive abilities', also at a level of aggregation that is more insightful. For this characterisation of abilities we look for an intermediate level of detail, excluding very specific abilities and skills (e.g., music skills, mathematical skills, hand dexterity, driving, etc.) but also excluding very general abilities or traits that would influence all the others (general intelligence, creativity, etc.). As we just cover cognitive abilities, we also exclude personality traits (e.g., the big five (Fiske 1949)). Although we consider the latter essential for humans, their ranges can be simulated in machines by changing goals and objective functions.

For our purposes we use 14 categories as the result of the integration of several tables and figures from (Hernández-Orallo 2017b), originally collected from psychometrics, comparative psychology, cognitive science and artificial intelligence (see Figure 1). The 14 categories are defined as follow: *Memory processes* (MP), *Sensorimotor interaction* (SI), *Visual processing* (VP), *Auditory processing* (AP), *Attention and search* (AS), *Planning and sequential decision-making and acting* (PA), *Comprehension and compositional expression* (CE), *Communication* (CO), *Emotion and self-control* (EC), *Navigation* (NV), *Conceptualisation, learning and abstraction* (CL), *Quantitative and logical reasoning* (QL), *Mind modelling and social interaction* (MS), and *Metacognition and confidence assessment* (MC). The hierarchical theories of intelligence in psychology, animal cognition and the textbooks in AI are generally consistent (at least partially) with this list of abilities, or in more general and simple terms, with this way of organising the vast space of cognition. The definition of the cognitive abilities can be found in (Vold and Hernandez-Orallo 2019).

Mapping

To generate the mapping between labour-related tasks and cognitive abilities, a multidisciplinary group of researchers conducted an annotation exercise for each item of the task database. More precisely, in a cross-tabulation of the vector of tasks $\mathbf{t}$ of length $p = |\mathbf{t}| = 59$ and cognitive abilities a of length $m = |\mathbf{a}| = 14$, each annotator was asked to put a 1 in a task-ability correspondence matrix $\mathbf{W}$ (59×14) if an ability is inherently required, i.e. absolutely necessary to perform the respective task (see the rubric in the Appendix). In order to increase robustness in the annotations, we followed a *Delphi Method* approach (Dalkey and Helmer 1963), repeating this process in order to increase agreement among annotators, and finally obtaining the share in percentage terms for each combination of task and ability. Similarly, we also linked the cognitive abilities with our list of AI benchmarks (which will be also described in detail in the following sections). Specifically, a group of AI-specialised researchers was asked to consider how each AI benchmark

is related to each cognitive ability: in a cross-tabulation of the vector of benchmarks b of length $n = |\mathbf{b}| = 328$ and cognitive abilities a of length $m = |\mathbf{a}| = 14$, we put a 1 in the ability-benchmark correspondence matrix $\mathbf{R}$ (14×328) if an ability is inherently required, i.e. absolutely necessary to solve the respective benchmark. Full information about this mapping procedure can be found in (anonymous).

Two-way interpretation

We can then translate the benchmark intensity vector $\mathbf{b}$ to cognitive abilities as a matrix-vector multiplication $\mathbf{Rb} \rightarrow \mathbf{a}$ thus obtaining an ability intensity vector $\mathbf{a}$ (14×1). We can also analyse task intensity, by weighting the task-ability mapping matrix by the ability intensity vector $\mathbf{a}$ as a matrix-vector multiplication $\mathbf{Wa} \rightarrow \mathbf{t}$ thus obtaining a new task intensity vector $\mathbf{t}$ (59×1).

This gives us a leftward interpretation of Figure 1 as:

$$\mathbf{Rb} \rightarrow \mathbf{a} \text{ and } \mathbf{Wa} \rightarrow \mathbf{t}$$

which together makes $\mathbf{WRb} \rightarrow \mathbf{t}$. This is interpreted as "benchmarks require abilities, which are required for tasks".

By using this framework we can analyse flows in both directions mathematically. Therefore, we can also give the rightward interpretation as:

$$\mathbf{t}^\top\mathbf{W} \rightarrow \mathbf{a}^\top \text{ and } \mathbf{a}^\top\mathbf{R} \rightarrow \mathbf{b}^\top$$

which together makes $\mathbf{t}^\top\mathbf{WR} \rightarrow \mathbf{b}^\top$.This is interpreted as "tasks require abilities, which are required for benchmarks".

Note that since both $\mathbf{W}$ and $\mathbf{R}$ mean "requires" (in the direction of abilities), it makes sense to distribute the values when a task or a benchmark requires many abilities. So, assuming that more abilities require more effort, we normalise both $\mathbf{W}$ and $\mathbf{R}$ through abilities. This means that in $\mathbf{W}$ rows are normalised to sum up 1, and in $\mathbf{R}$ columns are normalised to sum up 1, and values are thus in $[0, 1]$.

Analysis and Results

Now we are ready to analyse the correspondence between the two edges of our model. By comparing the values of $\mathbf{b}$ as propagated rightwards from $\mathbf{t}$ ($\mathbf{t}^\top\mathbf{WR} \rightarrow \mathbf{b}^\top$) against the values of $\mathbf{b}$ that originate directly from the benchmark intensities, we see very low correlations between these vectors. Figure 9 in the appendix shows some discrepancy scatterplots illustrating this. This picture is general, and we can conclude that the intensities do not match: the focus on AI benchmarks today does not correspond with the labour activities having highest intensity according to our data. Could this be different? In order to answer this question, in what follows we will analyse the results bidirectionally, exploring several hypotheses and professional profiles.

From AI to labour

As an illustrative example of how the model can be used in a single direction, we can obtain the task intensity vector $\mathbf{t}$ from the original benchmark intensity vector $\mathbf{b}$. This illustrates the leftward interpretation of Figure 1.

While in Figure 7 in the Appendix we show how our model works when specific AI benchmarks are selected.

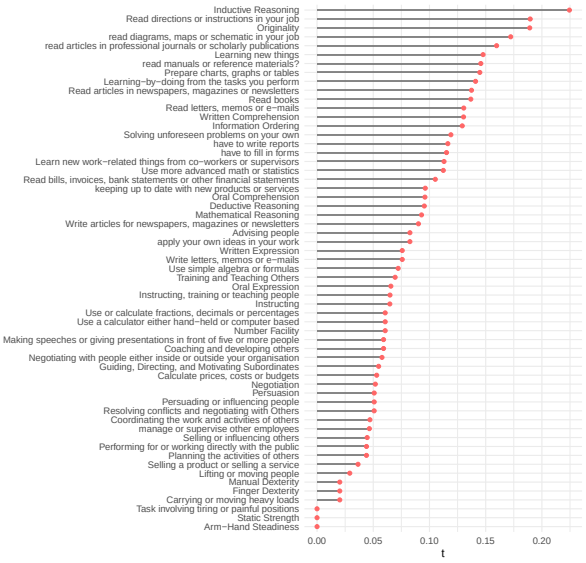

Figure 3: Labour-related tasks ranked in descending order based on by their intensity vector $\mathbf{t}$.

Figure 3 shows a sorted list of labour tasks according to the computed values in $\mathbf{t}$ from the analysis of *AI topics*. Those with the highest values consist almost entirely of information gathering and processing tasks (e.g., read letters or manuals, articles, bills, etc.), as well as performing tasks without using explicit instructions, relying on patterns and inference instead (e.g., learning, solving unforeseen problems, learning-by-doing, etc.). On the other hand, the lowest-scoring tasks are largely non-cognitive tasks that require a high degree of physical effort and dexterity (e.g., steadiness, manual/finger dexterity, etc.). This probably reflects a limited coverage of robotic benchmarks, which usually involve more propriosensory perception and manipulation. At the same time, there are also plenty of interpersonal tasks that include a human component. These are considered non-routine tasks (e.g., persuasion, supervision, communication or people management, etc.), all of which generally require social and emotional skills.

Note that the above considers the current activity (as extracted from the AAAI topics data) and the tasks that would be affected *if this activity would be transformed into progress in the areas the benchmarks represent and assuming that different abilities can be combined seamlessly*.

From labour to AI

Following the rightward interpretation of our setting, we can also analyse, given a particular (set of) occupation(s) and their corresponding set of tasks, which sort of AI benchmarks should attract more interest or require more effort from the AI research community in order to have a potential impact in the selected occupation(s).

We can do (1) one specific labour-related task or (2) a combination of tasks conforming particular occupations. Figure 8 in the Appendix shows some illustrative examples of (1). Regarding (2), we can also compute the AI benchmarks intensity scores by selecting six relevant occupations

from the ISCO-3 specifications: (a) general office clerks; (b) shop salespersons; (c) agricultural, forestry and fishery labourers; (d) medical doctors; (e) mining and construction labourers; (f) sales, marketing and public relations professionals; (g) mobile plant operators; (h) waiters and bartenders; (i) market gardeners and crop growers.

Because of the large number of AI benchmarks (328), we have clustered these benchmarks into six groups to make the interpretation of results easier (details in the appendix). Figure 4 depicts benchmark intensity scores for the nine selected occupations mentioned above. For instance, in order for AI developments to have an effect on general office clerks, AI research should focus on those benchmarks related to inspection and data extraction as well as on those focused on the development of narratives, question answering and social interaction.

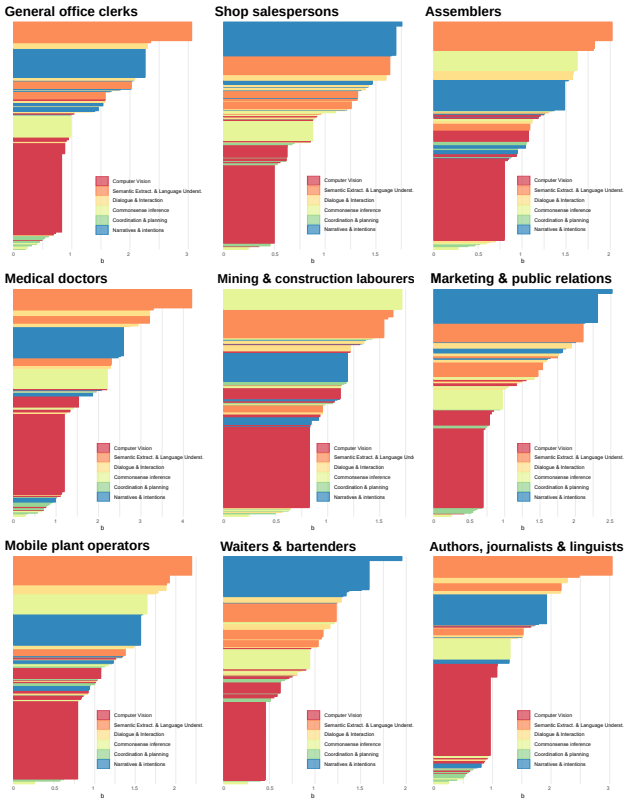

Figure 4: AI benchmarks ranked in descending order based on by their intensity vector $\mathbf{b}$ given their task intensity vectors $\mathbf{t}$ from six different occupations. Benchmarks coloured according to the cluster they belong to.

If we pay attention to those benchmarks where more progress is apparently taking place in AI (visual and auditory perception using deep learning and sensorimotor interaction, through (deep) reinforcement learning), we see that these cognitive abilities are generally at the bottom for the six selected occupations. This means either that (1) some of these skills are taken from granted (e.g., recognising objects and moving around in the workplace) or (2) many tasks in the workplace require skills for which there is not a high AI research activity at the moment. About (1), in our annotations, we included abilities when 'absolutely necessary'. Consequently, we considered that many of the tasks used in the workplace do not inherently require that a robot or a human visually recognises static or moving elements, as other capabilities could be used instead (e.g., blind people may "read manuals or reference manuals" using Braille).

Conclusions

We have developed a setting for the analysis of the relationship between Artificial Intelligence and the labour market in both directions. The setting combines occupations and tasks from the labour market with AI research benchmarks through an intermediate layer of cognitive abilities. The identification of the specific cognitive abilities that can be performed by AI gives a broader understanding on the impact of AI, as the inner layer is more independent of particular occupations, tasks or AI benchmarks. Although not included in the paper, we can also generate simulations outwards, setting a particular combination of ability intensities and propagate how tasks and occupations would be affected and what benchmarks would be more relevant. This analysis could also be done inwards.

In the paper we have seen examples where we can assess, in a very detailed way, how technological intensity of AI research may affect work-related tasks and corresponding occupations, as well as the other way round: how task and occupation intensity should be translated into AI research. We have seen the discrepancy between AI intensity and labour intensity and have used this setting to rank tasks by potential AI impact. In the end, we can determine which areas of AI research should be intensified if we sought to have a technological impact in particular selected task and occupations.

Despite its popularity in AI, using AI benchmarks to pulse the progress of AI research is fraught with caveats and criticisms, especially if performance metrics are used as an indication of progress. Instead, our model is based on intensities: we analyse whether some located activity on one edge translates on some located activity on the other edge. We use proxies for activities (such as time spent in a particular labour-related task or the research activity as per Figure 2). The use of activity versus progress makes this setting adoptable for the governance and assessment of AI R&D in academia and industry. In future work this analysis can be refined as more data becomes available on the relevance of specific work-related tasks as well as new AI benchmarks are introduced. Overall, we already present a powerful and flexible open tool[9] to map AI research and the impact on labour bidirectionally. The major merit of our model is not being predictive, but being prescriptive: we can decide priorities and make AI research interventions accordingly, to

[9]The presented setting and posterior analysis is flexible by updating data about benchmarks and professions, as well as the computed rates of intensity in AI benchmarks as measured using *AI topics*. Further details about the complete set of occupations, tasks, benchmarks and the associated intensity rates based on the results from *AI topics* or work surveys can be found in https://safe-tools.dsic.upv.es/shiny/OTAAI/.

procure that AI does qualify for the job.

Acknowledgments

This material is based upon work supported by the EU (FEDER), and the Spanish MINECO under grant RTI2018-094403-B-C3, the Generalitat Valenciana PROMETEO/2019/098. F. Martínez-Plumed was also supported by INCIBE (Ayudas para la excelencia de los equipos de investigación avanzada en ciberseguridad), the European Commission (JRC) HUMAINT project (CT-EX2018D335821-101), and UPV (PAID-06-18). J. H-Orallo is also funded by an FLI grant RFP2-152.

References

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

Acemoglu, D.; Dorn, D.; Hanson, G. H.; Price, B.; et al. 2014. Return of the Solow paradox? IT, productivity, and employment in us manufacturing. *American Economic Review* 104(5):394–99.

Arntz, M.; Gregory, T.; and Zierahn, U. 2016. The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis. *OECD Social, Employment and Migration Working Papers* 2(189):47–54.

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

Autor, D. 2013. The "task approach" to labor markets: an overview. Technical report, National Bureau of Economic Research.

Borg, I., and Groenen, P. 2003. Modern multidimensional scaling: Theory and applications. *Journal of Educational Measurement* 40(3):277–280.

Brynjolfsson, E.; Mitchell, T.; and Rock, D. 2018. What can machines learn, and what does it mean for occupations and the economy? In *AEA Papers and Proceedings*, volume 108, 43–47.

Buchanan, B. G.; Eckroth, J.; and Smith, R. 2013. A virtual archive for the history of ai. *AI Magazine* 34(2):86–86.

Dalkey, N., and Helmer, O. 1963. An experimental application of the delphi method to the use of experts. *Management science* 9(3):458–467.

Doddington, G. R.; Mitchell, A.; Przybocki, M. A.; Ramshaw, L. A.; Strassel, S. M.; and Weischedel, R. M. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, 1. Lisbon.

Eckersley, P.; Nasser, Y.; et al. 2017. EFF AI progress measurement project.

Felten, E. W.; Raj, M.; and Seamans, R. 2018. A method to link advances in artificial intelligence to occupational abilities. In *AEA Papers and Proceedings*, volume 108, 54–57.

Fernández-Macías, E., and Bisello, M. 2017. Measuring The Content and Methods of Work: a Comprehensive Task Framework. Technical report, European Foundation for the Improvement of Living and Working Conditions.

Fernández-Macías, E.; Bisello, M.; Sarkar, S.; and Torrejón, S. 2016. Methodology of the construction of task indices for the European Jobs Monitor. Technical report, European Foundation for the Improvement of Living and Working Conditions.

Fernández-Macías, E.; Gómez, E.; Hernández-Orallo, J.; Loe, B. S.; Martens, B.; Martínez-Plumed, F.; and Tolan, S. 2018. A multidisciplinary task-based perspective for evaluating the impact of AI autonomy and generality on the future of work. *CoRR* abs/1807.02416.

Fiske, D. W. 1949. Consistency of the factorial structures of personality ratings from different sources. *The Journal of Abnormal and Social Psychology* 44(3):329.

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

Goos, M.; Manning, A.; and Salomons, A. 2009. Job polarization in europe. *American economic review* 99(2):58–63.

Hernández-Orallo, J. 2017a. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review* 48(3):397–447.

Hernández-Orallo, J. 2017b. *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press.

Kodinariya, T. M., and Makwana, P. R. 2013. Review on determining number of cluster in k-means clustering. *International Journal* 1(6):90–95.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28(2):129–137.

Martinez-Plumed, F., and Hernandez-Orallo, J. 2018. Dual indicators to analyse ai benchmarks: Difficulty, discrimination, ability and generality. *IEEE Transactions on Games* 1–1.

Martínez-Plumed, F.; Avin, S.; Brundage, M.; Dafoe, A.; hÉigeartaigh, S. Ó.; and Hernández-Orallo, J. 2018. Accounting for the neglected dimensions of ai progress. *arXiv preprint arXiv:1806.00610*.

Nedelkoska, L., and Quintini, G. 2018. OECD Social, Employment and Migration Working Papers No. 38. Technical Report 38.

Shoham, Y.; Perrault, R.; Brynjolfsson, E.; Clark, J.; Manyika, J.; Niebles, J. C.; Lyons, T.; Etchemendy, J.; and Bauer, Z. 2018. The ai index 2018 annual report.

Vold, K., and Hernandez-Orallo, J. 2019. Ai extenders: The ethical and societal implications of humans cognitively extended by ai. In *AAAI / ACM Conference on Artificial Intelligence, Ethics, and Society*.

Wellman, M. P.; Greenwald, A.; Stone, P.; and Wurman, P. R. 2003. The 2001 trading agent competition. *Electronic Markets* 13(1):4–12.

This appendix contains supplementary material that is not strictly needed to follow the paper but adds more details about the procedures, methods and more illustrative examples of the use of our model. After acceptance, when the anonymity is resolved, an extended version of this appendix is found along with the information about the data and the tool.

Cognitive Abilities Rubric

We integrate several seminal psychometric models of intelligence to construct the following rubric of cognitive abilities.

Memory processes (MP)

Part of the information that is processed is stored in an appropriate medium to be recovered at will according to some keys, queries or mnemonics. This covers long-term memory and episodic memory, possibly using external devices such as books, spreadsheets, logs, databases, annotations, agendas and any other kind of analogical or digital recording and retrieval of data.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human stores new memories to be recovered at a future time?
- *Note*: the ability is about creating new memories, not only recovering them. We exclude short-term and working memory, as almost any cognitive task requires them.

Sensorimotor interaction (SI)

This deals with the perception of things, recognising patterns in different ways and manipulating them in physical or virtual environments with parts of the body (limbs) or other physical or virtual actuators, not only through various sensory and actuator modalities but in terms of mixing representations.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human perceives the surrounding physical or virtual world, the body and the manipulation of objects with the physical properties of these objects?
- *Note*: this may be done through different modalities, e.g., blind people can do this well or a bat/robot using a radar.

Visual processing (VP)

This deals with the processing of visual information, recognising objects and symbols in images and videos, movement and content in the image, with robustness to noise and different angles and transformations.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human recognises static or moving elements in images or videos?
- *Note*: this processing excludes the assessment of the consistence of what is seen.

Auditory processing (AP)

This deals with the processing of auditory information, such as speech and music, in noise environments and at different frequencies.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human recognises specific sounds, signals, alarms, speech, melodies, rhythm, etc.?
- *Note*: in the case of speech, we exclude the full understanding of sentences or the subjective perception of harmony in music.

Attention and search (AS)

This deals with focusing attention on the relevant parts of a stream of information in any kind of modality, by ignoring irrelevant objects, parts, patterns, etc. Similarly, it is the ability of seeking those elements that meet some criteria in the incoming information.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human identifies, tracks or focuses on elements that meet some criteria, especially when surrounded by other elements not meeting the criteria?
- *Note*: criteria may be about any perceptual modality, and they can also be categories: for instance, focusing on the trajectory of straws in a stream of water or instruments in a symphony.

Planning and sequential decision-making and acting (PA)

This deals with anticipating the consequences of actions, understanding causality and calculating the best course of actions given a situation.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human evaluates the effects of different sequences of events, plan various courses of actions and make a decision accordingly?
- *Note*: this excludes complex reasoning processes about the world and assumes planning under mostly consistent information. Note also that we are not referring to simple actions or decisions, as almost any cognitive system makes actions; the task must involve sequences, time or other dependencies to be considered under planning.

Comprehension and compositional expression (CE)

This deals with understanding natural language, other kinds of semantic representations in different modalities, extracting or summarising their meaning, as well as generating and expressing ideas, stories and positions.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human understands text, stories and other representations of ideas in different formats, and the composition or transformation of similar texts, stories or narratives, summarising or expressing ideas?
- *Note*: this may be done through different modalities: text, auditory, drawings, etc. Note also that we are not referring to the processing of simple and predefined phrases or symbols; the task must involve the understanding or compositional use of elements that make a whole: sentences, stories, summaries, etc..

Communication (CO)

This deals with exchanging information with peers, understanding what the content of the message must be in order to obtain a given effect, following different protocols and channels of informal and formal communication.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human communicates information between peers or units, using different kinds of protocols and channels, at different registers, ensuring that the messages are sent, received and processed appropriately by all the interested peers?
- *Note*: this excludes the narratives that the messages may contain, focusing on the effective channels of information.

Emotion and self-control (EC)

This deals with understanding the emotions of other agents, how they affect their behaviour and also recognising the own emotions and controlling them and other basic impulses depending on the situation.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human understands emotions of others/themselves, when they are true or fake, expressing the right emotional reactions, controlling and using them in the appropriate context?
- *Note*: this excludes the complexities of social modelling and anticipation.

Navigation (NV)

This deals with being able to move objects or oneself between different positions, through appropriate, safe routes and in the presence of other objects or agents, and changes in the routes.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human transfers objects and oneself from one place to another at different scales (rooms, buildings, towns, landscape, roads, etc.), using basic concepts for locations and directions?
- *Note*: this may be done through different modalities, and approaches such as landmarking, geolocations, etc..

Conceptualisation, learning and abstraction (CL)

This deals with being able to generalise from examples, receive instructions, learn from demonstrations, and accumulate knowledge at different levels of abstraction.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human generate different levels of abstractions, provided by peers or self-generated, acquiring knowledge incrementally built upon previously acquired knowledge?
- *Note*: this ability to learn or to abstract must be present and happen to complete the task; in other words, the task is not limited to the use of abstractions or concepts or operations learnt in the past.

Quantitative and logical reasoning (QL)

This deals with the representation of quantitative or logical information that is intrinsic to the task, and the inference of new information from them that solves the task, including probabilities, counterfactuals and other kinds of analytical reasoning.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human produces new conclusions or facts from quantities, logical facts or rules given as inputs, detecting inconsistencies and fallacies?
- *Note*: this goes beyond the simple combination of rules or instructions, such as ordering a deck of cards. Note also that we are not referring to the internal processing of symbols or numbers that are not part of the task, such as the potentials of a neuron, the instructions of a programming language or the arithmetic of a CPU/GPU.

Mind modelling and social interaction (MS)

This deals with the creation of models of other agents, so that their beliefs, desires and intentions can be understood, and anticipate the actions and interests of other agents.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human successfully interacts in social contexts with other agents having beliefs, desires and intentions, the understanding of group dynamics, leadership and coordination?
- *Note*: this is not about sociability or agreeableness, i.e., how willing an agent is to social situations.

Metacognition and confidence assessment (MC)

This deals with the evaluation of the own capabilities, reliability and limitations, self-assessing the probability of success, the effort and risks of own actions.

- *Rubric question*: Do all instances of this task inherently require that a robot or a human recognises accurately their own capabilities and limitations, when to assume responsibilities and when to delegate tasks and risks according to competences?
- *Note*: this goes beyond those cases covered by planning when considering the outcomes of several actions or no action. Note also that we are not referring to the mere selection of the action with highest probability or utility, as this is necessary for almost any task. This ability is about estimating and using the confidence of actions appropriately.

Cluster analysis of AI benchmarks

We performed a cluster analysis to simplify the analysis of intensities of the 328 AI benchmarks. We used the underlying structure of their required cognitive abilities. In this regard, we applied a k-means algorithm (Lloyd 1982), deciding the number of clusters k according to the elbow method (Kodinariya and Makwana 2013). This procedure minimises the total within-cluster variance up to the point where adding an additional cluster does not increase the percentage of

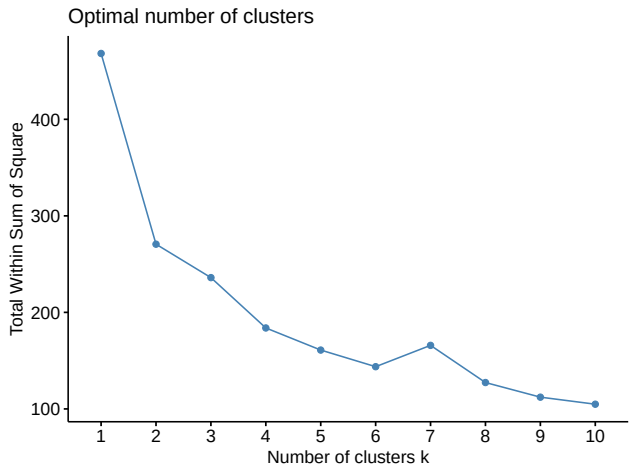

Figure 5: Elbow criterion reached in 6 groups when clustering AI benchmarks given the underlying structure of their required cognitive abilities.

variance explained. Figure 5 shows the results of the elbow method, where $k = 6$ groups seems to be a good choice.

In order to gain intuitive understanding of the the regularity governing the relationships among the selected 6 clusters of AI benchmarks, Figure 6 shows their projection on a three-dimensional cube identified by the three principal dimensions of a multidimensional scaling procedure (Borg and Groenen 2003). This procedure creates an optimal low-dimensional configuration of the original (multidimensional) data creating a map displaying the relative positions of a number of objects, given only a table of the distances between them.

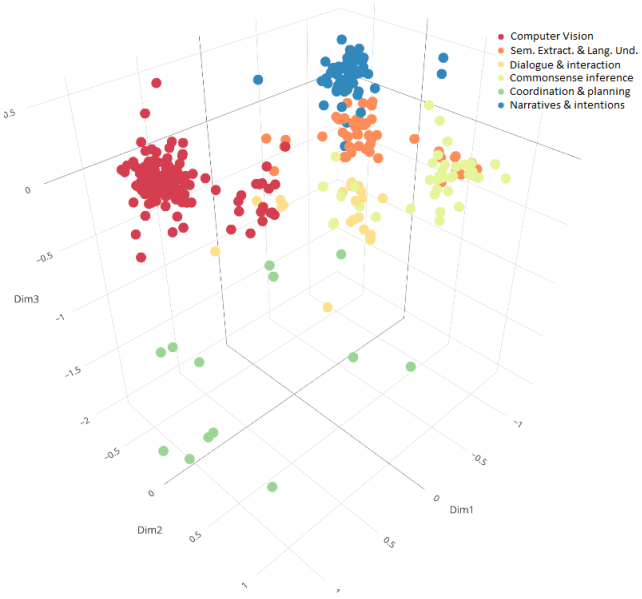

Figure 6: Three-dimensional scaling of $\mathbf{R}$. Points are coloured according to the cluster they belong to.

- **Cluster 1** (*Computer Vision*): This cluster can be characterised mostly with computer vision-related benchmarks. Some examples of benchmarks in this cluster are MNIST, ImageNet, Pascal3D, CIFAR or COCO.
- **Cluster 2** (*Semantic Extraction and Language Understanding*): This cluster includes some tasks dealing with information extraction using Natural Language Processing. Some examples of benchmarks in this cluster are CoNLL, ACE, LexNorm, Yelp Dataset or the Stanford Natural Language Inference (SLNI) Corpus.
- **Cluster 3** (*Dialogue and interaction*): This cluster groups benchmarks that are related to interaction (between humans and machines), testing dialogue and speech performance. This cluster includes benchmarks such as Wizard-of-Oz dataset, Loebner Prize, other variants of the Turing Test or the Robochat challenge.
- **Cluster 4** (*Commonsense inference*): This cluster includes tasks related to learning and handling commonsense knowledge, such as data mining, knowledge bases, reasoning and commonsense, recommendation, etc. Some examples of benchmarks in this cluster are UCI, FB15k, Winograd Schema Challenge, Event2Mind or MovieLens.
- **Cluster 5** (*Coordination and planning*): This cluster includes games and different multi-agent benchmarks, including planning, coordination, collaboration, etc. Examples of benchmarks in this cluster are ALE, GVGAI, Robocup, RLComp, Go or Angry Birds.
- **Cluster 6** (*Narratives and intentions*): This cluster is characterised by narratives, question answering, sentiment analysis and other reading comprehension tasks. Examples of benchmarks in this cluster are SQuAD, Quora Question Pairs, QAngaroo, SemEval or SentEval.

From AI benchmarks to labour-related tasks: illustrative examples

Following the leftward interpretation of our setting (e.g., $\mathbf{WRb} \rightarrow \mathbf{t}$), we can analyse which labour tasks would be affected if we aimed at emphasising one specific AI benchmark.

In Figure 7 we can see a couple of illustrative examples: (top) shows that negotiation, coordination, planning, guiding and other persuasion-related tasks are intensified if the *Trading Agent Competition* (TAC) (Wellman et al. 2003), the benchmark challenge for competing AI agents, is set as the focus in AI research; (bottom) shows that written and reading communication tasks and activities are intensified if the *Automatic Content Extraction* program (Doddington et al. 2004), a benchmark for entities, relations, and the events recognition in text, is the focus in AI research.

From labour-related tasks to AI: illustrative examples

Following the rightward interpretation of our setting (e.g., $\mathbf{t}^\top\mathbf{WR} \rightarrow \mathbf{b}^\top$), we can analyse which AI benchmarks would require more effort if we aimed at emphasising one specific labour-related task.

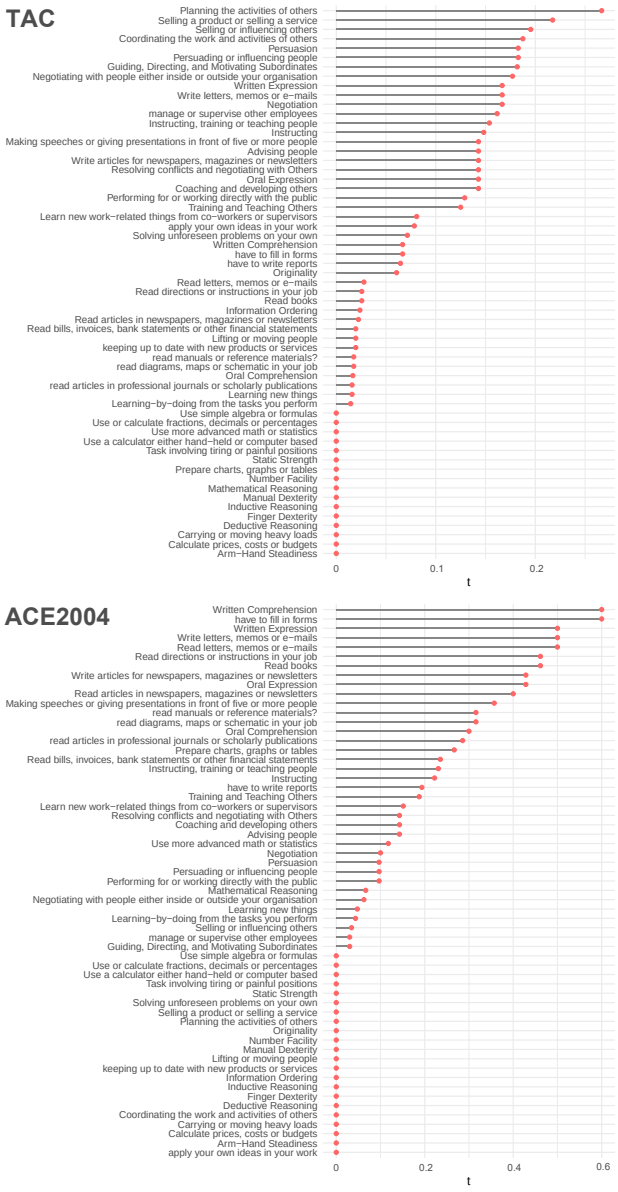

Figure 7: Labour-related tasks ranked in descending order based on by their intensity vector $\mathbf{t}$ where a sigle benchmark is selected, using the intensities coming from AI topics: (top) Trading Agent Competition (TAC) (Wellman et al. 2003) (bottom) Automatic Content Extraction (ACE) benchmark (Doddington et al. 2004)

In Figure 8 we can see a some illustrative examples: (top-left) in order to have a potential effect on the "*instructing" task* the focus of AI research should be put on AI benchmarks related to interacting and dynamic scenarios for autonomous software agents testing coordination and planning as well those related to semantic extraction and natural language understanding should be the focus in AI research; (top-right) in order to have a potential effect on the "*Lifting or moving people*" task the focus of AI research should be put on AI benchmarks related to

planning and coordination multi-agent scenarios and, to a much lesser extent, to computer vision; (bottom-left) in order to have a potential effect on the "*coordination*" task the focus of AI research should be put on AI benchmarks related to dialogue and interaction (between humans and machines) benchmarks as well as those related to coordination and planning in multi-agent systems to be intensified; finally, (bottom-right) in order to have a potential effect on the "*Solving unforeseen problems on your own*" the focus of AI research should be put on AI benchmarks related to commonsense inference and computer vision.

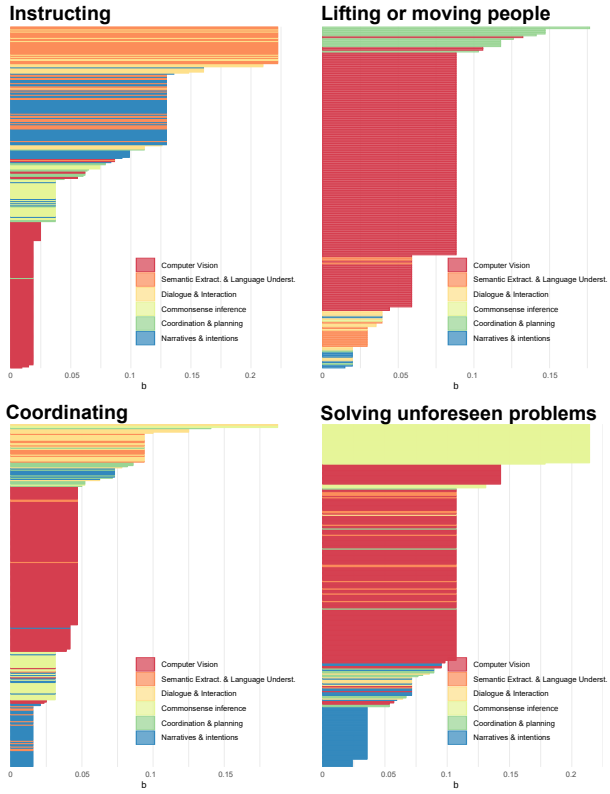

Figure 8: AI benchmarks ranked in descending order conditional on by their intensity vector $\mathbf{b}$ given a task intensity vector $\mathbf{t}$. Plots show those AI benchmarks that should be intensified when we focus on specific (set of) labour-related tasks: (top-left) "*Advising people*"; (top-right) "*Lifting or moving people*"; (bottom-left) "*Coordinating*"; (bottom-right) "*Solving unforeseen problems on your own*".

Further details about this and other examples, as well as the complete description of the set of occupations, tasks, benchmarks and the associated intensity rates based on the results from *AI topics* or work surveys can be found in https://safe-tools.dsic.upv.es/shiny/OTAAI/

Discrepancy between AI and labour intensities

In this section we analyse whether the current intensity in labour and AI match for those analysed occupations in Figure 4. In order to check this, in Figure 9 we show discrep-

ancy scatterplots in which we compare the intensity vector $\mathbf{b}$ obtained from *AI topics* (as explained in section "AI benchmarks") with the intensity vector $\mathbf{b}$ obtained using the rightward interpretation of our setting (e.g., $\mathbf{t}^\top\mathbf{WR} \rightarrow \mathbf{b}^\top$) when we emphasise one specific occupation.

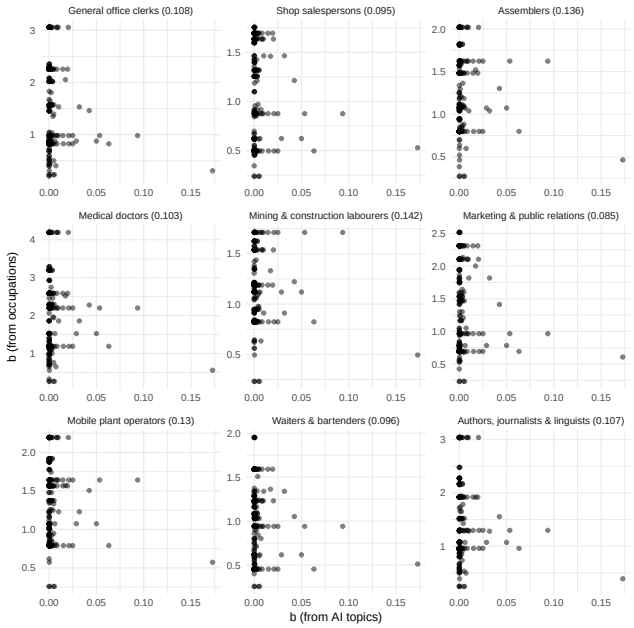

Figure 9: Discrepancy scatterplots between the intensity vector $\mathbf{b}$ obtained from *AI topics* and those obtained using the rightward interpretation of our setting. Values in parentheses (in the titles) show the Spearman correlations.

We can see that the different intensity vectors obtained per occupation do not match current intensity in AI for any occupations in the figure and, in general, for any of all the set of 119 occupations we are analysing in our setting. Figure 9 also show that the Spearman correlations are close to 0, so there is no rank correlation between the different intensity vectors, meaning that the those tasks that present high intensity in the workplace do not correspond to those benchmarks presenting high activity.

Therefore, the answer for the question *does AI qualify for the job?* is *not yet*.