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

Fernando Martínez-Plumed
Universitat Politècnica de València
fmartinez@dsic.upv.es

Songül Tolan
Joint Research Centre
European Commission
songul.tolan@ec.europa.eu

Annarosa Pesole
Joint Research Centre
European Commission
annarosa.pesole@ec.europa.eu

José Hernández-Orallo
Universitat Politècnica de València
jorallo@upv.es

Enrique Fernández-Macías
Joint Research Centre
European Commission
enrique.fernandez-macias@ec.europa.eu

Emilia Gómez
Joint Research Centre
European Commission
emilia.gomez-gutierrez@ec.europa.eu

APPENDIX

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.

A COGNITIVE ABILITIES RUBRIC

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

A.1 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.

A.2 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.

A.3 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.

A.4 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.

A.5 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.

A.6 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.

A.7 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..

A.8 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.

A.9 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.

A.10 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..

A.11 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.

A.12 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.

A.13 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.

A.14 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.

B 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 [4], deciding the number of clusters k according to the elbow method [3]. This procedure minimises the total within-cluster variance up to the point where adding an additional cluster does not increase the percentage of variance explained. Figure 1 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 2 shows their projection on a three-dimensional cube identified by the three principal dimensions of a multidimensional scaling procedure [1]. This procedure creates an optimal low-dimensional configuration of the original (multi-dimensional) data creating a map displaying the relative positions of a number of objects, given only a table of the distances between them.

- **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

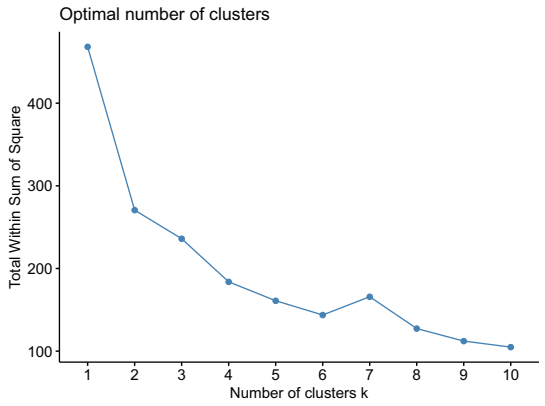

Figure 1: Elbow criterion reached in 6 groups when clustering AI benchmarks given the underlying structure of their required cognitive abilities.

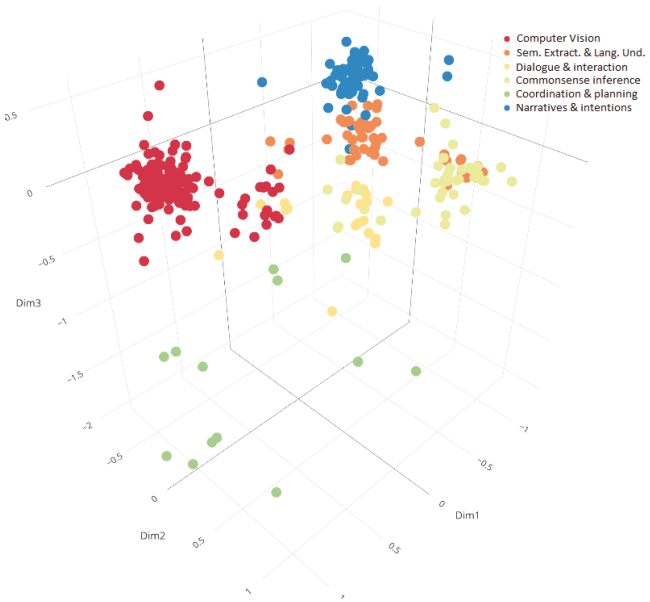

Figure 2: Three-dimensional scaling of R. Points are coloured according to the cluster they belong to.

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.

C 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 3 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) [5], 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 [2], a benchmark for entities, relations, and the events recognition in text, is the focus in AI research.

D 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.

In Figure 4 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.

Further details about this and other examples, as well as the complete description of the set of occupations, tasks, benchmarks

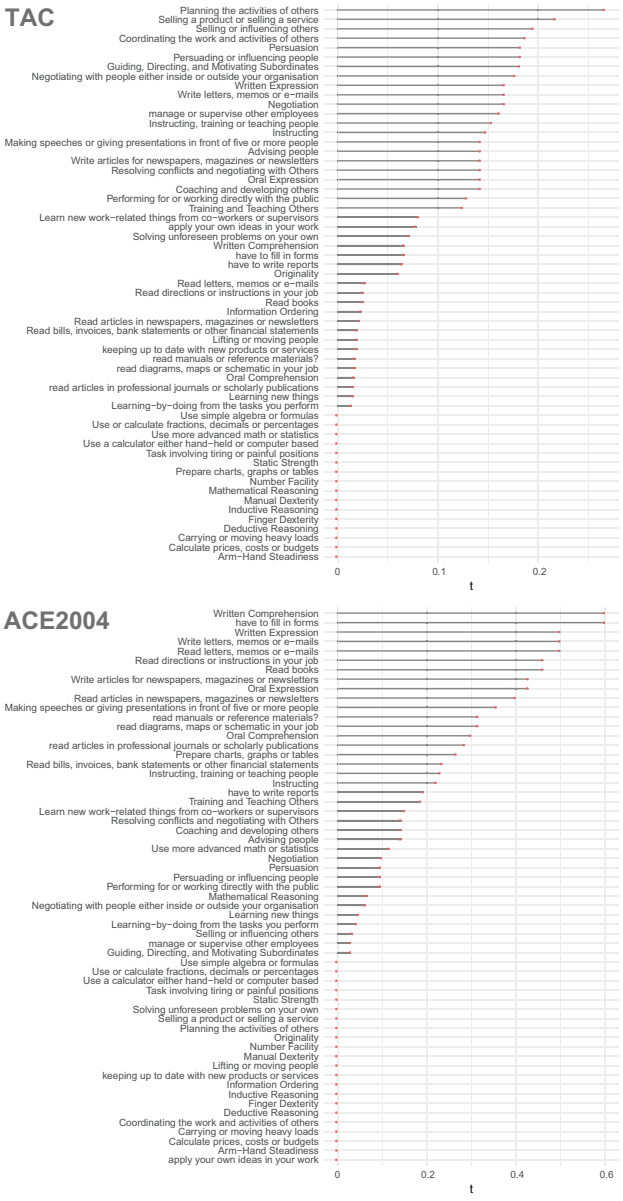

Figure 3: Labour-related tasks ranked in descending order based on by their intensity vector t where a sigle benchmark is selected, using the intensities coming from AI topics: (top) Trading Agent Competition (TAC) [5] (bottom) Automatic Content Extraction (ACE) benchmark [2]

and the associated intensity rates based on the results from *AI topics* or work surveys can be found in *link anonymised*

E 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 **??**. In order to check this, in Figure 5 we show discrepancy scatterplots in

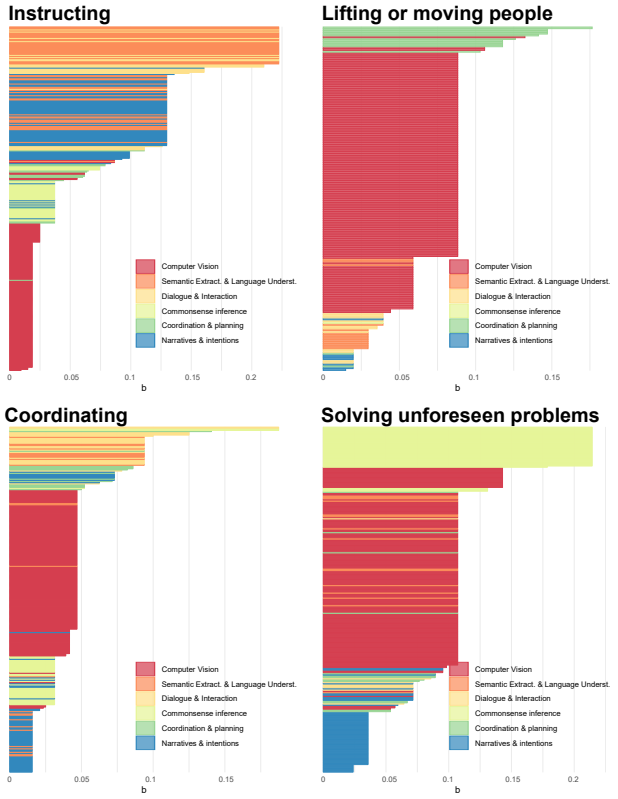

Figure 4: AI benchmarks ranked in descending order conditional on by their intensity vector b given a task intensity vector t. Plots show those AI benchmarks that should be intensified when we focus on specific (set of) labour-related tasks: (top-left) "*Advising people*"; (top-right) "*Lifting or moving people*"; (bottom-left) "*Coordinating*"; (bottom-right) "*Solving unforeseen problems on your own*".

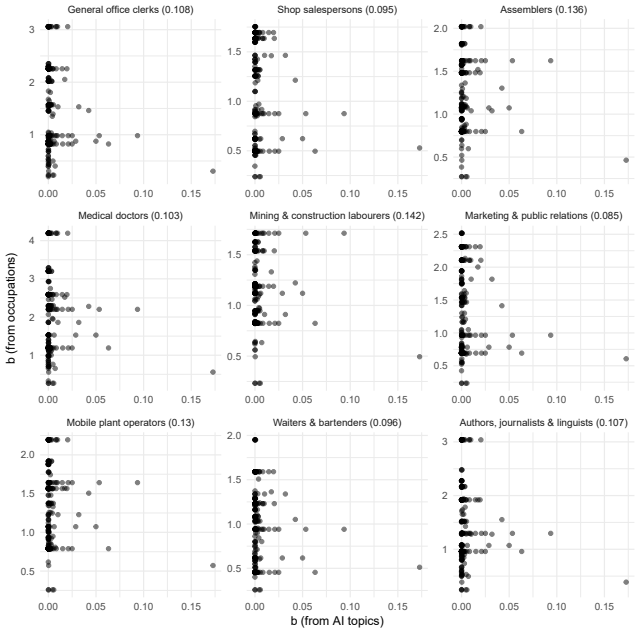

Figure 5: Discrepancy scatterplots between the intensity vector b obtained from *AI topics* and those obtained using the rightward interpretation of our setting. Values in parentheses (in the titles) show the Spearman correlations.

which we compare the intensity vector $\mathbf{b}$ obtained from *AI topics* (as explained in section "**??**") 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.

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 5 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.*

REFERENCES

[1] Ingwer Borg and Patrick Groenen. 2003. Modern multidimensional scaling: Theory and applications. *Journal of Educational Measurement* 40, 3 (2003), 277–280.

[2] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation.. In *Lrec*, Vol. 2. Lisbon, 1.

[3] Trupti M Kodinariya and Prashant R Makwana. 2013. Review on determining number of Cluster in K-Means Clustering. *International Journal* 1, 6 (2013), 90–95.

[4] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28, 2 (1982), 129–137.

[5] Michael P Wellman, Amy Greenwald, Peter Stone, and Peter R Wurman. 2003. The 2001 trading agent competition. *Electronic Markets* 13, 1 (2003), 4–12.