Journal of Artificial Intelligence Research 71 (2021) 191-236 Submitted 01/2021; published 06/2021

# Measuring the occupational impact of AI: tasks, cognitive abilities and AI benchmarks

**Songül Tolan** SONGUL.TOLAN@EC.EUROPA.EU
**Annarosa Pesole** ANNAROSA.PESOLE@EC.EUROPA.EU
**Fernando Martínez-Plumed** FERNANDO.MARTINEZ-PLUMED@EC.EUROPA.EU
**Enrique Fernández-Macías** ENRIQUE.FERNANDEZ-MACIAS@EC.EUROPA.EU
*Joint Research Centre, European Commission*

**José Hernández-Orallo** JORALLO@UPV.ES
*Universitat Politècnica de València*
*Leverhulme Centre for the Future of Intelligence*

**Emilia Gómez** EMILIA.GOMEZ-GUTIERREZ@EC.EUROPA.EU
*Joint Research Centre, European Commission*
*Universitat Pompeu Fabra*

## Abstract

In this paper we develop a framework for analysing the impact of Artificial Intelligence (AI) on occupations. This framework maps 59 generic tasks from worker surveys and an occupational database to 14 cognitive abilities (that we extract from the cognitive science literature) and these to a comprehensive list of 328 AI benchmarks used to evaluate research intensity across a broad range of different AI areas. The use of cognitive abilities as an intermediate layer, instead of mapping work tasks to AI benchmarks directly, allows for an identification of potential AI exposure for tasks for which AI applications have not been explicitly created. An application of our framework to occupational databases gives insights into the abilities through which AI is most likely to affect jobs and allows for a ranking of occupations with respect to AI exposure. Moreover, we show that some jobs that were not known to be affected by previous waves of automation may now be subject to higher AI exposure. Finally, we find that some of the abilities where AI research is currently very intense are linked to tasks with comparatively limited labour input in the labour markets of advanced economies (e.g., visual and auditory processing using deep learning, and sensorimotor interaction through (deep) reinforcement learning).

## 1. Introduction

There is wide agreement that the latest advances in Artificial Intelligence (AI), driven by rapid progress in machine learning (ML) and its subfields, will have disruptive repercussions on the labour market (Shoham et al., 2018). Previous waves of technological progress have also had a sustained impact on labour markets (Autor and Dorn, 2013), yet the notion prevails that the impact of ML will be different (Brynjolfsson et al., 2018). An argument that supports this notion is that ML seems to circumvent the previously hard limit to automation known as Polanyi's Paradox (Polanyi, 1966), which states that we humans "know more than we can tell". While past technologies could only automate tasks that follow explicit, codifiable rules, ML technologies can infer rules automatically from the observation of inputs and corresponding outputs (Autor,

2014). This implies that ML may facilitate the automation of many more types of tasks than were affected in previous waves of technological progress (Brynjolfsson et al., 2018).

Our perception of what AI is able to do is driven by the growing importance of benchmarks in AI (Hernández-Orallo et al., 2017). For instance, a decisive moment for deep learning really happened when it started to perform better than many other techniques in benchmarks such as ImageNet (Deng et al., 2009) and CIFAR10 (Krizhevsky et al., 2009). The rhythm is so hectic that new benchmarks appear everyday and replace old ones, setting the bar higher and higher.[1] In the end, breakthroughs in some particular challenges and benchmarks have been identified as landmarks of the field (Campbell et al., 2002; Ferrucci, 2012; Mnih et al., 2015; Silver et al., 2016; Brown and Sandholm, 2019; Brown et al., 2020) and used as illustrations of what AI can do. Also, the activity around benchmarks is a good indicator of where the research effort in AI is focusing.

In this paper we develop a framework for analysing the potential occupational impact of AI (illustrated in Figure 1). [2] The explicit focus on AI distinguishes this analysis from studies on robotisation (Acemoglu and Restrepo, 2018), digitalisation and online platforms (Agrawal et al., 2015), and the general occupational impact of technological progress and automation (Autor, 2015). That is, automation through technologies that do not require AI, e.g. self-checkout machines that replace human cashiers in supermarkets, is not considered in this framework. The framework links tasks to cognitive abilities, and these to indicators that measure performance in different AI fields. More precisely, we map 59 generic tasks from the worker surveys European Working Conditions Survey (EWCS) and Survey of Adult Skills (PIAAC) as well as the occupational database O*Net to 14 cognitive abilities (that we extract from the cognitive science literature) and these to a comprehensive list of 328 AI evaluation tasks from benchmarking initiatives, challenges, competitions and scientific literature. These AI-related metrics reflect the intensity of current research and development in different AI techniques. This "research intensity" indicator is not necessarily a good proxy of future AI progress, since breakthroughs do not always appear where more research effort is spent, and there may be dead ends that are not obvious yet. But future AI progress is simply impossible to predict, and we believe our approach provides a sensible approximation to where AI may have a bigger impact in the short and medium term, since we directly measure where more research effort is spent (see Figure 1).

Differently from previous approaches that directly link AI developments with task characteristics (Brynjolfsson et al., 2018), our framework adds an intermediate layer of cognitive abilities. With 14 distinct cognitive abilities, this layer is more detailed than the task characteristics mentioned in the task-based approach by Autor et al. (2003). In this earlier model work tasks are defined by their routine, abstract, and manual content, all three characteristics of work that point towards task automation (Autor and Handel, 2013). Although this approach has been very fruitful and inspired many studies (including this one), in our view these characteristics do not suffice to capture AI's potential to affect and transform work tasks that are not (yet) tailored to be per-

1. For instance, CIFAR10 was followed by the more challenging CIFAR100 (Krizhevsky et al., 2009), SQuAD1.1 has been replaced by SQuAD2.0 (Rajpurkar et al., 2018), GLUE by SUPERGLUE (Wang et al., 2019), Starcraft by Starcraft II (Vinyals et al., 2017) and the new Arcade Learning Environment (ALE) (Machado et al., 2018) by the PlayStation Reinforcement Learning Environment (PSXLE) (Purves et al., 2019).
2. All the data, code and results can be found in `https://github.com/nandomp/AIlabour`

formed (fully or partially) by a machine. Hence, in this paper we try to identify what kinds of task content (and occupations) are more likely to be impacted by AI advances currently in the making, without assuming that such an impact implies labour substitution. As argued by Bessen (2019) among others, in the past, new productive technologies have much more often transformed than replaced occupations, and this is also the most likely effect of current AI advances on the future of work, at least in the short and medium term. Additionally, the technical feasibility of automatically performing a given type of task content is not a sufficient condition for a large-scale substitution of human by machines for that content, as other factors such as the relative cost of labour, work organisation and the elasticity of demand also have to be taken into account (Autor, 2013; Fernández-Macías et al., 2018; Bessen, 2019)

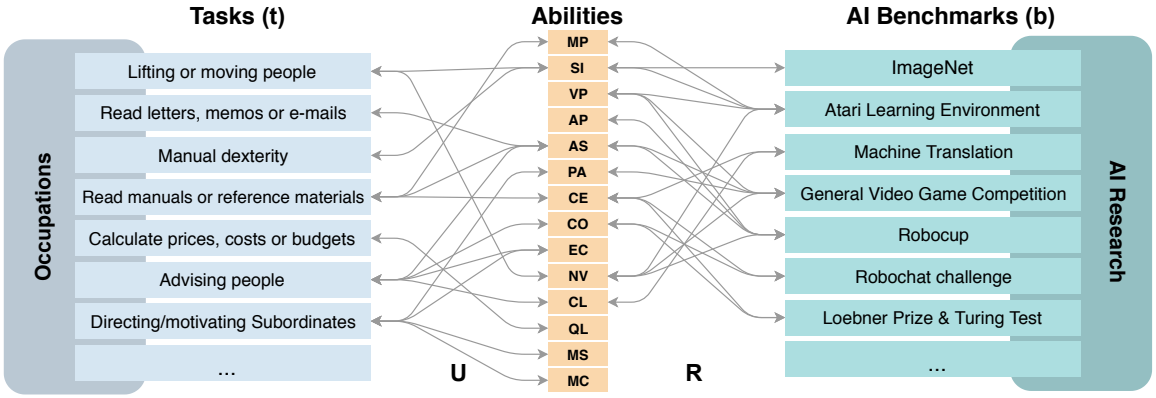



Figure 1: Illustrative example of the bidirectional and indirect mapping between job market and Artificial Intelligence (abilities described in Appendix A). The full list of tasks is presented in Table 6 of Appendix C. The full list of AI benchmarks is presented in Table 7 of Appendix D.The notation we use is **t** for the tasks, **a** for the abilities and **b** for the benchmarks. The arrows are represented by correspondence matrices **U** (task-ability correspondence) and **R** (ability-benchmark correspondence).

The ability perspective allows us to distinguish machines that, through AI, are empowered with the abilities of performing a range of several tasks from machines that are explicitly constructed or programmed to perform specific tasks. For instance, the ability of understanding human language (Manning et al., 1999) can be applied in a variety of tasks (such as reading or writing e-mails, or advising costumers/clients). Abilities are therefore a better parameter to evaluate progress in AI (Hernández-Orallo, 2017a). Note that general abilities are different from skills: from a human perspective abilities are innate and primary. Instead, skills are acquired through a combination of abilities, experience and knowledge for some specific domain, but could be obtuse for other problems (Fernández-Macías et al., 2018). Here, we focus on abilities instead of skills. Since knowledge and experience are not suitable properties of AI, linking AI benchmarks to abilities (instead of skills) should be less prone to measurement error (Hernández-Orallo, 2017a).

Due to the intermediate layer of 14 different abilities, we also gain a broader understanding on the occupational exposure to AI. That is, the framework allows not only to define a single occupation-level AI exposure score, but also to identify the specific tasks affected and the different abilities that are most likely driving the implementation of AI in the workplace. Conversely,

we can identify which abilities are less likely to be performed by AI and are therefore less prone to changes in the way they are currently being performed.

Furthermore, we rely on a wide range of AI benchmarks to approximate the direction of AI research and development. These benchmarks are linked to performance metrics (such as classification accuracy, squared error, perplexity, AUC, etc.) on openly accessible datasets. These are prominently placed either in the scientific literature or on online platforms where both AI researchers and industry players present their current performance in different AI domains. The collection of these benchmarks provides a thorough overview of the direction of AI progress. In many cases these benchmarks and the work on them precede the explicit formalisation of its use at work. For instance, performing well in games such as checkers, chess, Go and poker (Silver et al., 2016, 2017; Brown and Sandholm, 2019), which is recorded in corresponding benchmarks, is not a required ability in any work-related task. However, AI that performs well on these benchmarks needs to exhibit abilities in memory processing, learning and planning. These abilities are useful in the performance of some work-related tasks.

In addition, connecting these benchmarks to work-related tasks allows to explore the question of the occupational impact of AI in the other direction, from occupational needs to specific AI benchmarks. That is, following the framework illustrated in Figure 1, we can identify occupations that are less exposed to AI, filter out the tasks that need to be performed in these occupations and specify which of the required abilities can be connected to corresponding benchmarks (and benchmark clusters) that would require increases in research activity for AI to have an impact on these occupations.

Instead of looking at past progress of these benchmarks, we measure interest in AI domains through the presence of benchmarks in each category. This allows for the computation of future trends based on past developments in each category and can be easily updated for future years. This repository of AI benchmarks is open and accessible[3] (Martínez-Plumed et al., 2020a,b).

The remainder of this paper is structured as follows. The next section embeds the present study into the literature, which is followed by background information on the construction of the layer tasks and cognitive abilities in the framework. After presenting in Section 4 the methodology used to construct the framework, we describe in Section 5 the different data sources that we combined to construct the framework. We present the results of the application of our framework in Section 6, and discuss them in Section 7. Section 8 concludes.

## 2. Related Work

This study complements the work done by Martínez-Plumed et al. (2020), which introduces the framework that connects AI benchmarks and work tasks through cognitive abilities for the first time. That work presents how the framework can be used to analyse the relationship between AI and occupations in a bidirectional way (from AI to occupations, from occupational requirements to AI) and descriptive results with a list of tasks (see Table 6, Appendix C) and clustered AI benchmarks (see Table 7, Appendix D). The focus of that work were AI benchmarks and the

3. http://www.aicollaboratory.org/

prescriptive implications of the framework. In this work, we rather focus on the occupational impact, and use a more refined methodology to measure work contents. More specifically, we index the task impact by the task framework presented in Table 1. This re-balances the task impact as measured by Martínez-Plumed et al. (2020) (which is more driven by the availability of task data, than by a conceptual framework) according to an established model of work for a more accurate measure of work contents. In addition, this paper provides insights on the economic background of the impact of AI on the labour market by discussing the findings in the context of wages, technology driven labour-market polarisation and AI diffusion. Finally, this paper discusses the validation of the framework and provides a comparison of some of the main findings to other relevant results from the economic literature.

This paper contributes to the literature on the occupational impact of recent technological change (Frey and Osborne, 2017; Arntz et al., 2016; Nedelkoska and Quintini, 2018), although, in contrast with most of these studies, we do not try to estimate the extent of (potential) labour replacement. Instead, we aim at identifying which occupations and types of task contents are more directly related to current developments in AI research, and therefore are more likely to be affected by applications of AI to work in the future. According to some more recent literature (Brynjolfsson et al., 2018), AI may lead to substitution effects. However, the more likely effect of AI on work in many cases is complementarity (AI as a tool). Some of our findings point in that direction. However, this is not something we explicitly discuss in this paper. This approach captures the entire AI research field more comprehensively than expert predictions on the future automatability of occupations as by Frey and Osborne (2017) and subsequent studies.

This measure of AI progress complements the rubric by Brynjolfsson et al. (2018) used to determine the suitability of tasks for ML since it can be easily updated to future developments in the already recorded benchmarks. In addition, some of the task properties listed in the rubric may be endogenous to the redefinition of an occupational task for which AI applications have explicitly been constructed. For instance, the property "large (digital) data sets exist or can be created containing input-output pairs" is only a task property once the task is explicitly considered for AI. By contrast, with the ability perspective we identify the potential impact of AI on a task, by looking at the abilities that both AI and humans need to possess in order to perform such task, with no need of further defining new properties at the sole aim of integrating AI.

A different approach to rubrics is the use of expert feedback. For instance, this has been done to assess the time frame for the so-called human-level machine intelligence (HLMI), the point where AI would outsmart humans (Müller and Bostrom, 2014, 2016). A more specific take has been explored by Grace et al. (2018), which not only includes questions about HLMI or automating "all professions", but also a series of particular professions or tasks, such as surgeon, truck driver, New York Times best-seller writer, human-level language translator, retail salesperson, Atari games player, Starcraft player, laundry folder or champion of the world series of poker. The list is not comprehensive but gives a very informative view of how diverse the predictions are depending on the activity. For instance, experts estimate that AI surgeons will achieve human performance in about 35 years from 2018, whereas doing fold laundry automatically as well as a human is estimated to happen in only about 6 years from 2018. Expert polls and forecasting using Delphi or other consensus methods are powerful ways to estimate when some milestones

will happen, complementary to the methodology we use in this paper. Using benchmark activity, apart from being methodologically very different, allows to do a more granular and systematic analysis and obtain a perception of research intensity in AI.

Our approach relates most to Felten et al. (2018), who also link AI field benchmarks to work-related abilities, but there are some noteworthy differences. First, Felten et al. (2018)measure AI progress on one particular platform, the Electronic Frontier Foundation (EFF)[4], which is restricted to a more limited set of AI benchmarks. The benchmarks in the present framework further rely on our own previous analysis and annotation of papers (Hernández-Orallo, 2017b; Martínez-Plumed et al., 2018; Martinez-Plumed and Hernandez-Orallo, 2018) as well as on open resources such as *Papers With Code*[5], which include data and results from a comprehensive set of AI benchmarks, challenges, competitions and tasks. This ensures a broad coverage of AI tasks, also providing insight into AI performance in cognitive abilities that go beyond perception, such as language processing, planning, information retrieval or automated deduction/induction.

For better comparability across these benchmarks that come from many different AI domains, the measure of AI intensity is also different. Felten et al. (2018) assess AI progress by computing linear trends in each benchmark. However, nonlinear performance jumps at different thresholds of progression (i.e. breakthroughs) of each benchmark impede comparability between them. We address this by translating benchmarks to AI research activity, which we consider more comparable across benchmarks from different AI fields.

Finally, Webb (2020) measures the occupational impact of AI by computing the overlap between O*NET job task descriptions and the text of patents. AI-related patents are identified based on matches with general AI keywords. We complement this approach and offer a broader picture by mapping research intensity in specific AI domains (e.g., computer vision or natural language processing) to particular abilities required to perform job tasks.

## 3. Background

Before giving details on the methodology on how we develop the framework, we provide some background information on the literature from which we draw the concepts for tasks and cognitive abilities and elaborate on them in the following sections.

### 3.1 Tasks

In our framework (see Figure 1), occupations are decomposed into a vector of tasks. Tasks have been defined as units of work activity that produce output (Autor, 2013). From our perspective, that derives from Fernandez-Macias and Bisello (2020), each occupational task can be understood as a specific act of transformation on an object. On the basis of the type of object being transformed and the type of transformation, we can create a taxonomy of different types of tasks. At the highest level, this classification differentiates between tasks that operate on material

4. https://www.eff.org/es/ai/metrics
5. https://paperswithcode.com/

*things* (physical tasks), tasks that operate on *ideas* or information (intellectual tasks) and tasks that operate on social relations with *people* (social tasks). From those, a nested taxonomy with increasing levels of detail unfolds, as shown in Table 1. In this paper, we use 59 indicators at the most detailed level of the tasks taxonomy shown in Table 6.[6]

**Content**

1. **Physical tasks**
   - (a) Strength
   - (b) Dexterity

2. **Intellectual tasks**
   - (a) Information processing:
     - (I) I.P. of uncodified information
     - (II) I.P. of codified information
       - (i) Literacy:
         - (a) Business
         - (b) Technical
         - (c) Humanities
       - (ii) Numeracy:
         - (a) Accounting
         - (b) Analytic
   - (b) Problem solving:
     - (I) Information gathering and evaluation.
     - (II) Creativity and resolution.

3. **Social tasks**
   - (a) Serving/attending
   - (b) Teaching/training/ coaching
   - (c) Selling/influencing
   - (d) Managing/ coordinating

**Methods and tools**

1. **Work organisation**
   - (a) Autonomy
   - (b) Teamwork
   - (c) Routine
     - (I) Repetitiveness
     - (II) Standardization

2. **Technology**
   - (a) Machines (excluding ICT)
   - (b) Information and Communication technologies
     - (I) Basic ICT
     - (II) Programming

Table 1: A classification of tasks according to their contents and methods (source: Fernandez-Macias and Bisello (2020))

.

### 3.2 Cognitive Abilities

A first glance at the tasks that are usually identified in the workplace and those that are usually set in AI as benchmarks (see a sample in Figure 1) reveals the difficulty of matching them directly, as the lists are very different. However, tasks and benchmarks have some latent factors in common, what we refer to as "cognitive abilities", which we can use to map them indirectly but 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 a car, 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): openness, conscientiousness, extraversion, agreeableness and neuroticism). Although we consider the latter essential for humans, their ranges can be simulated in machines by changing goals and objective functions.

At the intermediate level, we aim at a number and breadth similar to the "broad abilities" of the Cattell-Horn-Carroll hierarchical model (see Figure 2) (Carroll et al., 1993; Keith and Reynolds,

6. Note that the task framework by Fernandez-Macias and Bisello (2020) has been updated in May 2020. In this paper we refer to an earlier version of the framework that does not include some task indices that have been added to the left-hand side (Content) of Table 1. However, the differences are very minor.

2010). However, some of them are very anthropocentric and not distinctive enough to sufficiently cover all aspects of cognition.

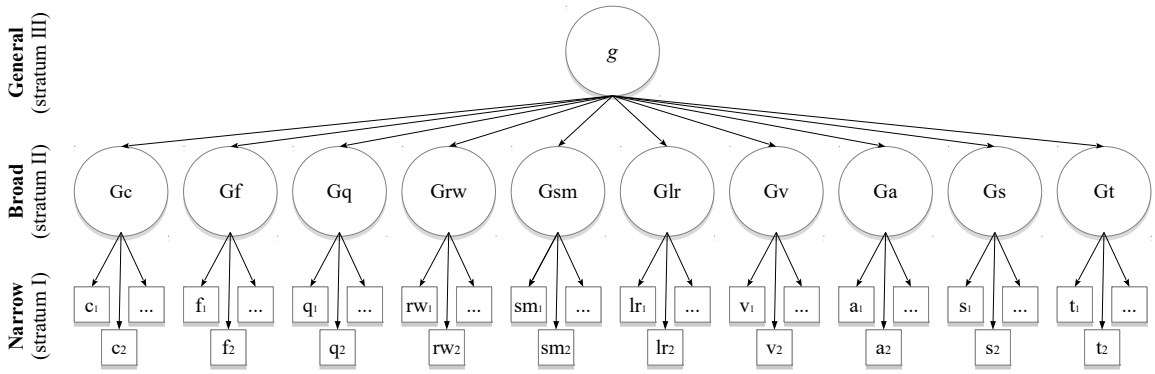



Figure 2: Cattell-Horn-Carroll's three stratum model. The broad abilities are Crystallised Intelligence (Gc), Fluid Intelligence (Gf), Quantitative Reasoning (Gq), Reading and Writing Ability (Grw), Short-Term Memory (Gsm), Long-Term Storage and Retrieval (Glr), Visual Processing (Gv), Auditory Processing (Ga), Processing Speed (Gs) and Decision/Reaction Time/Speed (Gt).

For our purposes we use a taxonomy of 14 cognitive abilities, merging several categorisations in psychology, animal cognition and AI, originally introduced by Hernández-Orallo and Vold (2019). The elements that compose this taxonomy were distilled from different sources. In particular, the list draws from Thurstone's primary mental abilities, according to (Schaie, 2010), the factors from Cattell-Horn-Carroll hierarchical model (Carroll et al., 1993; Keith and Reynolds, 2010), (stratum II, Figure 2), the areas of animal cognition research according to the table of contents of (Wasserman and Zentall, 2006), the main areas in AI according to the AI Journal (as per 2017), the "competency" areas in AGI according to (Adams et al., 2012) and the I-athlon "events" from (Adams et al., 2016). These different lists were merged into an integrated list by matching synonyms and related terms, and trying to keep a manageable number of broad capabilities. There were tensions between both distinctiveness and comprehensiveness against the number of abilities. The main criterion for keeping a distinction between two abilities A and B (and not merging them) was the understanding that a system or component (either natural or artificial) could conceivably master one of them while failing at the other. The compromise for completeness was easier to find; some elements (such as processing or decision speed in the Cattell-Horn-Carroll) are not proper abilities; also, some abilities related to multimodality were not explicitly included in the final list of 14 (e.g., olfactory processing). The current version only covers "visual" and "auditory" processing, being the two most representative sensory modalities.

The 14 categories are: Memory processes **(MP)**, Sensorimotor interaction **(SI)**, Visual processing **(VP)**, Auditory processing **(AP)**, Attention and search **(AS)**, Planning, sequential decision-making and acting **(PA)**, Comprehension and 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)**, 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 cognitive abilities can be found in Appendix A, which also includes a *rubric* so that we can determine for each ability whether it is required for a particular task.

## 4. Methodology

In this section we explain the construction of the framework. We map between the three layers: (1) tasks (2) cognitive abilities, and (3) AI research.

| Matrix | Description | Unit | Appears in Section |
|---|---|---|---|
| $\mathbf{t^o}$ $(59\times1)$ | tasks intensity vector for occupation $\mathbf{o}$ | $\forall\ i'th$ element of $\mathbf{t^o}, \mathbf{t_i} \in [0,1]$ | 4.1, 5.1 |
| $\Omega$ $(59\times14)$ | task-ability correspondence annotation matrix | $\forall \omega_{ij} \in \Omega, \omega_{ij} \in \{0,1,...,6\}$ | 4.1, B |
| $\mathbf{U}$ $(59\times14)$ | task-ability correspondence matrix | $\forall u_{ij} \in \mathbf{U}, u_{ij} \in \{0,1\}$ | 4.1, B |
| $\Psi^{\mathbf{o}}$ $(59\times14)$ | task intensity-ability matrix for occupation $\mathbf{o}$ | $\forall \psi_{ij} \in \Psi, \psi_{ij} \in [0,1]$ | 4.1 |
| $\Phi^{\mathbf{o}}$ $(14\times14)$ | task indices-ability matrix for occupation $\mathbf{o}$ | $\forall \phi_{ij} \in \Phi^{\mathbf{o}}, \phi_{ij} \in [0,1]$ | 4.1 |
| $\mathbf{W}$ $(119\times14)$ | ability intensity - occupation matrix | $\forall w_{ij} \in \mathbf{W}, w_{ij} \in [0,1]$ | 4.1,4.3,6.1 |
| $\mathbf{R}$ $(328\times14)$ | ability-benchmark correspondence matrix | $\forall r_{ij} \in \mathbf{R}, r_{ij} \in \{0,1\}$ | 4.2 |
| $\mathbf{b}$ $(328\times1)$ | benchmark intensity vector | $\forall\ i'th$ element of $\mathbf{b}, \mathbf{b_i} \in [0,1]$ | 4.2,5.2 |
| $\mathbf{a}$ $(14\times1)$ | ability-specific AI intensity vector | $\forall\ i'th$ element of $\mathbf{a}, \mathbf{a_i} \in [0,1]$ | 4.2,4.3,6.2 |
| $\mathbf{V}$ $(119\times14)$ | occupation-ability AI impact matrix | $\forall v_{ij} \in \mathbf{V}, v_{ij} \in [0,1]$ | 4.3,6.3 |

Table 2: Summary of notation.

Following the framework in Figure 1 from left to right, we construct an index for 119 standardised occupations, using information on each occupation's intensity of 59 tasks, which are in turn linked with 14 cognitive abilities. Then, these 59 links per occupation and cognitive ability are sorted into the 14 task indices (as shown on the left of Table 1). Going further right in Figure 1, we also link these 14 cognitive abilities to 328 AI benchmarks. In total, we construct the framework based on information for 119 standardised occupations, 59 tasks, 14 task indices, 14 cognitive abilities, and 328 AI benchmarks. We summarise the notation in Table 2. A more detailed explanation of this notation follows below.

### 4.1 Work Tasks to Cognitive Abilities

This section elaborates on the mapping between work tasks and cognitive abilities. To generate matrix $\mathbf{U}(59\times14)$ (see Table 4 for an excerpt), we conducted a manual annotation exercise in a

multidisciplinary[7] group of seven researchers of which six were annotators for each variable of the task database. More precisely, starting with matrix $\Omega(59 \times 14)$ (see Table 3 for an excerpt), the number of task variables (see Appendix C for the list of tasks) is the row dimension and the number of cognitive abilities (see Appendix A for list and definition of cognitive abilities) is the dimension in the columns. Each annotator was asked to put a 1 in a cell if an ability is inherently required, i.e., absolutely necessary to perform the respective task. In order to increase robustness in the annotations, we followed a *Delphi Method* approach (Dalkey and Helmer, 1963)[8]. In other words, manual annotations were conducted independently and iteratively in two rounds. In the second round the annotation exercise was repeated but in light of the results and discussions of the first round. To increase robustness in the assignment of abilities to tasks (and to obtain the curated Boolean matrix $\mathbf{U}(59 \times 14)$), we defined an ability as assigned to a task if at least two annotators assigned this ability. This makes the assignment less sensitive to outlier assignments of individual annotators. Thus $\mathbf{U}[\omega_{\mathbf{ij}} \geq \mathbf{2}]$.

Note that the annotations of the second round were neither random nor independent. Thus, matrix $\Omega$ is the result of an iterative process (the Delphi method) that is directed at converging to a unified matrix of annotations. We abbreviate this process by implementing the above mentioned rule ($\mathbf{U}[\omega_{\mathbf{ij}} \geq \mathbf{2}]$ ) to obtain matrix $\mathbf{U}$. However, to allow for uncertainty (i.e., to consider the fact that at each task-ability cell an annotator may have opted for a different annotation), we perturb the values of $\Omega$ by one standard deviation (per column) upwards and downwards.

| **Task** | **MP** | **SI** | **VP** | **AP** | **AS** | **PA** | **CE** | **CO** | **EC** | **NV** | **CL** | **QL** | **MS** | **MC** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Carrying or moving heavy loads | 0 | 6 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 |
| Standing | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Write letters, memos or e-mails | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| Calculate prices, costs or budgets | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 1 |
| Use more advanced maths or statistics | 2 | 0 | 0 | 0 | 6 | 0 | 2 | 0 | 0 | 0 | 1 | 6 | 0 | 0 |
| Resolving conflicts and negotiating | 1 | 0 | 0 | 0 | 1 | 3 | 5 | 6 | 6 | 0 | 1 | 2 | 6 | 4 |
| Instructing, training or teaching people | 4 | 0 | 0 | 0 | 1 | 1 | 6 | 6 | 1 | 0 | 1 | 0 | 5 | 1 |

Table 3: Excerpt of Matrix $\Omega$.

| **Task** | **MP** | **SI** | **VP** | **AP** | **AS** | **PA** | **CE** | **CO** | **EC** | **NV** | **CL** | **QL** | **MS** | **MC** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Carrying or moving heavy loads | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Standing | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Write letters, memos or e-mails | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Calculate prices, costs or budgets | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Use more advanced maths or statistics | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Resolving conflicts and negotiating | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| Instructing, training or teaching people | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |

Table 4: Excerpt of Matrix $\mathbf{U}$.

7. The disciplines represented were sociology and occupational research, cognitive psychology, computer science and economics.
8. See Appendix B for a more detailed description of the annotation exercise, including indicators on consensus and assignment behaviour among annotators.

To obtain matrices $\Psi^o$, the task intensity-ability matrices for every occupation **o**, we take the Hadamard product of every vector $\mathbf{t^o}$ with the correspondence matrix **U**:

$$\mathbf{U} \circ \mathbf{t^o} = (u_{i,j} \cdot t_i^o) = \begin{pmatrix} u_{1,1} \cdot t_1^o & \cdots & u_{1,14} \cdot t_1^o \\ \vdots & \ddots & \vdots \\ u_{59,1} \cdot t_{59}^o & \cdots & u_{59,14} \cdot t_{59}^o \end{pmatrix} = \Psi^o \tag{1}$$

Note that these 59 tasks are used to generate the left side of the task framework presented in Fernandez-Macias and Bisello (2020) that consists of 14 task indices. That is, different sets of $t^o$ contribute to the same concept of tasks which in turn require similar sets of abilities. To avoid that the ability scores are driven by data availability of tasks, we orthogonalise task information by averaging over task sets (rows of matrices $\Psi^{\mathbf{o}}(59 \times 14)$) that are assigned to the same task index. For each occupation **o** this reduces the rows of matrices $\Psi^{\mathbf{o}}(59 \times 14)$ from 59 (number of tasks) to 14 (number of task indices) which yields the task indices-ability matrix $\Phi^{\mathbf{o}}(14 \times 14)$.

In order to take into account the number of task indices that a cognitive ability is assigned to, we sum over all task indices linked to the same cognitive ability for each occupation. In addition, we take into account the additional complexity of combining multiple abilities in one occupation by normalising the ability-specific task intensities such that the sum of scores within each occupation is equal to one:

$$\frac{\sum\limits_i \phi_{\mathbf{i,j}}^{\mathbf{o}}}{\sum\limits_i \sum\limits_j \phi_{\mathbf{i,j}}^{\mathbf{o}}} = \mathbf{w^o} \tag{2}$$

Stacking each vector $\mathbf{w^o}$ yields matrix $\mathbf{W}(119 \times 14)$ which indicates the relative required intensity of each of the 14 cognitive abilities in each of the 119 occupations.

Note that the differences in the total intensities across different cognitive abilities are not linear, since the score of each cognitive ability derives from variables with highly varying scales. However, these scores take into account the number of tasks for which an ability is required weighted by the intensity of each task in each occupation. This allows for a ranking of the relevance of each ability within an occupation. Similarly, the scores for the same cognitive ability across different occupations are measured on the same scale, which allows for a ranking of occupations along the relevance of that cognitive ability for the occupation.

For the computation of an AI impact score, we want the ability-specific scores to take into account the additional energy that needs to be spent on coordination if multiple abilities need to be combined in one occupation at equally high intensity levels. That is, two very different occupations can have the same degree of intensity of one ability but can still be affected in very different ways by AI research intensity of this ability if the corresponding tasks require a different number of abilities at the same time. For instance, visual processing may be a very relevant ability for a person classifying offensive online content. Similarly, visual processing may be equally relevant for surgeons but also in combination with sensorimotor interaction. If we considered the intensity of each cognitive ability separately this would suggest that high AI intensity in visual processing but relatively low intensity in sensorimotor interaction would affect both occupations

equally. However, in reality the surgeon would be affected less than the person classifying online content because performing visual processing would have to be combined with strong sensorimotor interaction. Therefore, we employ the normalisation by each occupation in Equation 2. More specifically, this lowers the ability-specific scores for occupations with many high total ability-specific scores, such as medical doctors, but increases the ability-specific scores for occupations that only have high scores for one or two abilities.

### 4.2 AI Benchmarks to Cognitive Abilities

Similar to the mapping between cognitive abilities and tasks, we link these 14 cognitive abilities to the data on AI benchmarks (see Section 5.2). 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 $\mathbf{b}$ of length $|\mathbf{b}| = 328$ and the 14 cognitive abilities, a 1 is put in an ability-benchmark correspondence (or mapping) matrix $\mathbf{R}$ ($14 \times 328$) if an ability is inherently required, i.e., absolutely necessary to solve the respective benchmark.

From here we can calculate the vector of relevance for each cognitive ability from the correspondence matrix $\mathbf{R}$ as $\sum_j \mathbf{r}_{ij}$ as row. We normalise the relevance by the total number of documents to obtain the ability-specific AI intensity vector $\mathbf{a}$:

$$\frac{\sum\limits_{i} \mathbf{r_{i,j}}}{\sum\limits_{i}\sum\limits_{j} \mathbf{r_{i,j}}} = \mathbf{a} \tag{3}$$

### 4.3 Combining Occupations and AI Through Abilities

We combine AI benchmarks (see Section 4.2) to labour market information using the common link to cognitive abilities. For this purpose we take the Hadamard product of matrix $\mathbf{W}$ (see Section 4.1) with the respective AI research intensity vector $\mathbf{a}$:

$$\mathbf{W}^{\top} \circ \mathbf{a} = (w_{i,j} \cdot a_i) = \begin{pmatrix} w_{1,1} \cdot a_1 & \cdots & w_{1,119} \cdot a_1 \\ \vdots & \ddots & \vdots \\ w_{14,1} \cdot a_{14} & \cdots & w_{14,119} \cdot a_{14} \end{pmatrix} = \mathbf{V}^{\top} \tag{4}$$

We obtain a single AI exposure score for each occupation by taking the sum over the rows of matrix $\mathbf{V}$, i.e. $\sum_j \mathbf{v}_{ij}$. The final score indicates which of the studied occupations are relatively more likely to be affected by AI research intensity (i.e. which occupations are more exposed to AI progress) in the analysed cognitive abilities. For illustrative purposes we normalise this score, which we call AI exposure score, to a $[0, 1]$ scale.

## 5. Data

This section describes the data preparation process[9]. We rely on different sources of data that provide information on task intensity in occupations ($\mathbf{t^o}(59 \times 1)$ for each occupation $\mathbf{o}$), i.e., the

9. All the data, code and results can be found in `https://github.com/nandomp/AIlabour`

relevance of and time spent on that task, on the one side, and on AI research intensity ($\mathbf{b}(328 \times 1)$) on the other side, where neither side is based on one dataset only. We first describe how we generate the task dataset from the combination of existing occupational databases. Next, we describe the data gathering process deployed to generate the relevant AI benchmarks.

### 5.1 Tasks: Work Intensity

For the task dataset ($\mathbf{t^0}(59 \times 1)$), we draw from the framework developed by Fernandez-Macias and Bisello (2020). This data entails a list of tasks (presented in Table 6 in Appendix C) and their respective intensity (i.e., relevance and time spent) across occupations. In the following we provide a summary of the construction of this dataset.

We classify occupations according to the 3-digit International Standard Classification of Occupations (ISCO-3)[10]. Since there is no international data source that unifies information on all tasks required, we combine data from three different sources: two worker surveys: (1) the European Working Conditions Survey (EWCS)[11] and (2) the OECD Survey of Adult Skills (PIAAC)[12] as well as the Occupational Information Network (O*NET)[13].

The data in the worker surveys are measured at the individual worker level based on replies to questions on what they do at work. Task intensity is derived as a measure of time spent on specific tasks. For instance, in the EWCS we derive the task *"Lifting or moving people"* from the survey question $q24b$ *"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"*. Analogously, in the PIAAC we derive the task *"Read letters, memos or e-mails"* from the survey question G_Q01$b$ *"[In your main paid job] Do you read letters, memos or e-mails?"* and the corresponding 5-point scale answers ranging from *"Every day"* to *"Never"*. Due to the nature of survey data, we need to be aware of issues such as measurement error, high variation in responses across individuals and biased responses.

Similarly, the occupational database, O*NET is based on multiple waves of individual worker surveys but also on employer job postings, expert research and other sources. The data is curated by occupational experts and provided on a standardised occupational level. In this case, task intensity is derived from a variable that measures the extent to which the task is required to perform a job. For instance, the task *"Oral Comprehension"* is derived from the same variable and the corresponding level defined on a 7-point scale.

The O*NET is widely used in the literature on labour markets and technological change (Acemoglu and Autor, 2011; Frey and Osborne, 2017; Goos et al., 2009). Moreover, it covers a large share of the task list that we construct. However, the occupational level of the data precludes a further analysis into variation in task content within occupations. Moreover, much like the EWCS for Europe, the O*NET is based on US data only. Therefore, likely differences in the task content of occupations across countries due to institutional as well as socio-economic differences can-

10. https://www.ilo.org/public/english/bureau/stat/isco/
11. https://www.eurofound.europa.eu/surveys/european-working-conditions-surveys
12. https://www.oecd.org/skills/piaac/
13. https://www.onetonline.org/

not be considered in the present analysis[14].

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 which is equivalent to the task-intensity vectors $\mathbf{t^o}$ for each occupation $\mathbf{o}$.

To test the consistency of the variables that are derived from multiple datasources, Fernandez-Macias and Bisello (2020) look at pairwise correlations and Cronbach's Alpha for multiple variables that measure similar concepts. Reassuringly, all tests yield high correlations and Cronbach's Alpha values of between 0.8 and 0.9, suggesting consistency in the measurement of task intensity across the different data sources. However, it is reasonable to doubt the comprehensiveness of the task framework. In fact, continuing research on this topic has led to the addition of some further indicators in the task framework (more specifically the left side of Table 1). Since the collection of data on these additional task categories is yet to be conducted, the results of this paper will not include them. However, the impact of these in our results should be marginal[15].

### 5.2 Benchmarks: AI Intensity

For the present framework we generate a comprehensive repository of AI benchmarks (Martínez-Plumed et al., 2020a,b) based on our previous compilation, analysis and annotation of AI papers and benchmarking results (Hernández-Orallo, 2017a; Martínez-Plumed et al., 2018; Martinez-Plumed and Hernandez-Orallo, 2018; Martínez-Plumed et al., 2020a,b) as well as open resources such as *Papers With Code*[16] (the largest, up to date, free and open repository of machine learning code and results), which includes data from several AI-related repositories (e.g., EFF[17], NLP-progress[18], SQuAD[19], RedditSota[20], etc.). All these repositories draw on data from multiple (verified) sources, including academic literature, review articles and code platforms focused on machine learning and AI.

For the purposes of this study, from the aforementioned sources we track the reported evaluation results (when available or sufficient data is provided) on different metrics of AI performance across separate AI benchmarks (e.g., datasets, competitions, challenges, awards, etc.) from a

14. However, according to the analysis of (Fernández-Macías et al., 2016) the cross-country variation in task contents within occupations tends to be quite small. In fact, most of the variations are observed in work organisation and use of technology, which we do not consider in this paper.
15. The following task indices were added to the framework: 1.(Within Physical tasks) Navigation: moving objects or oneself in unstructured and changing spaces 2.(Within Intellectual - Information processing tasks) Visual and/or auditory processing of uncodified and unstructured information 3.(Within Intellectual - Problem Solving tasks) Information search and retrieval 4.(Within Intellectual - Problem Solving tasks) Planning
16. `https://paperswithcode.com/`
17. `https://www.eff.org/es/ai/metrics`
18. `https://github.com/sebastianruder/NLP-progress`
19. `https://rajpurkar.github.io/SQuAD-explorer/`
20. `https://github.com/RedditSota/state-of-the-art-result-for-machine-learning-problems`

number of AI domains, including (among others) 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, etc. This ensures a broad coverage of AI tasks, also providing insight into AI performance in cognitive abilities that go beyond perception, such as the ability to plan and perform actions on such plans. Note that most of these benchmarks we are addressing are *specific*, implying that their goals are clear and concise, and that researchers can focus on developing specialised AI systems for solving these tasks. This does not mean researchers are not allowed to use more general-purpose components and techniques to solve many of these problems, but it may be easier or most cost-effective for the researchers to build a strongly specialised system for the task at hand.

Our framework uses data from 328 different AI benchmarks for which there is enough information available to measure their progress for different evaluation metrics. Table 7 in Appendix D contains the details from the benchmarks used in our analysis.

When aiming at evaluating the progress in a specific (AI) discipline, we need to focus on objective evaluation tools to measure the elements and objects of study, assess the prototypes and artefacts that are being built and examine the discipline as a whole (Hernández-Orallo, 2017a). Depending on the discipline and task, there is usually a loose set of criteria about how a system is to be evaluated. See for instance Figure 3 showing the progress for various evaluation metrics of object recognition in the COCO benchmark (*Common Objects in COntext*) (Lin et al., 2014). Several questions might arise regarding the latter: How can we compare results or progress between different metrics? How to compare between different benchmarks for the same task (e.g., COCO vs. MNIST (Bottou et al., 1994) vs. ImageNet (Deng et al., 2009)) or different tasks for the same benchmark? Or, even more challenging, how can we compare results from different tasks in the same domain or different domains? Actually, although there might be a general perception of progress due to the increasing trends of the metrics (or decreasing in case of error-based measures), it would be misleading to consider that the progress in AI should be analysed by the progress of specific systems solving specific tasks, while there may be a complete lack of understanding of the relationships between different tasks. What does it mean, for instance, that one AI system demonstrates impressive (e.g., super-human) performance for a natural language processing task and another demonstrates impressive performance for a perception task (with respect to some evaluation metrics) if both developments cannot be integrated easily into a single agent in order to display more general perceptual or linguistic capabilities, at the same time (Brundage, 2016)? On the other hand, it is also hard to tell in many domains whether progress comes from better hardware, data, computing, software, and other resources, or better AI methods (Martínez-Plumed et al., 2018). Furthermore, the specialisation of many metrics to the domain, the evaluation overfitting (Whiteson et al., 2011), and the lack of continuity in some evaluation procedures can also be recognised as limitations and constraints (Hernández-Orallo, 2017a) when evaluating the progress of AI.

Given the above difficulties, instead of using the rate of progress, what we can analyse is the activity level around a specific benchmark, indicating the research intensity in a specific task in terms of the production (e.g., outputs such as research publications, news, blog-entries, etc.)

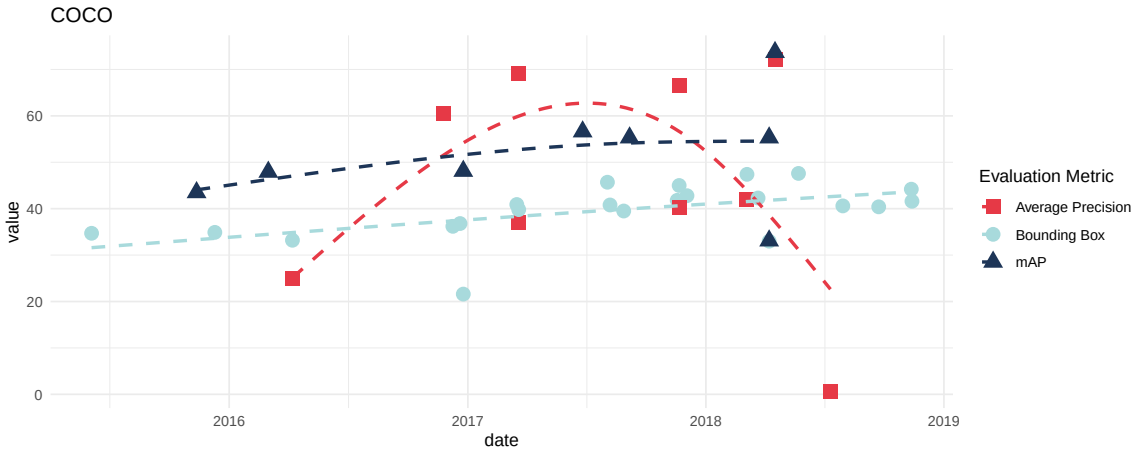



Figure 3: Progress (trends represented with dashed coloured lines) across different evaluation metrics for COCO object recognition benchmark (Krizhevsky et al., 2009).

from the AI community related to the above AI benchmarks. Benchmarks that have an increasing trend in their production rates 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. For instance, this has happened in areas such as machine translation and object recognition, where the research intensity has been very high in the past years and the progress and applications are undeniable. It is also worth considering that areas that usually gather more intensity are those where there is a general perception that breakthroughs are being made or about to be made. For instance, those problems that are already solved, where progress is expected to be minimal or those that are too challenging for the state of the art usually capture less attention.

We can derive the activity level or *intensity* using some proxies. In particular, we performed a quantitative analysis using data obtained from *AI topics*[21], an archive kept by the Association for the Advancement of Artificial Intelligence (AAAI)[22]. 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 et al., 2013). In this regard, in order to calculate the intensity for each particular benchmark, we average the normalised[23] number of hits (e.g., documents) obtained from *AI topics* per benchmark and year over a specific period of time (e.g, last year, lustrum or decade). This way we obtain the benchmark intensity vector $\mathbf{b}(328 \times 1)$ with values in $[0,1]$, as they are counts divided by the total number of documents. Figure 4 presents the calculated relative intensity for a set of illustrative AI benchmarks over the last decade. Note that we make the assumption that a high relative intensity corresponds to breakthroughs or significant progress that can be translated to real applications in the short term.

21. `https://aitopics.org`
22. `https://www.aaai.org/`
23. Document counts are normalised to sum up to 100% per year

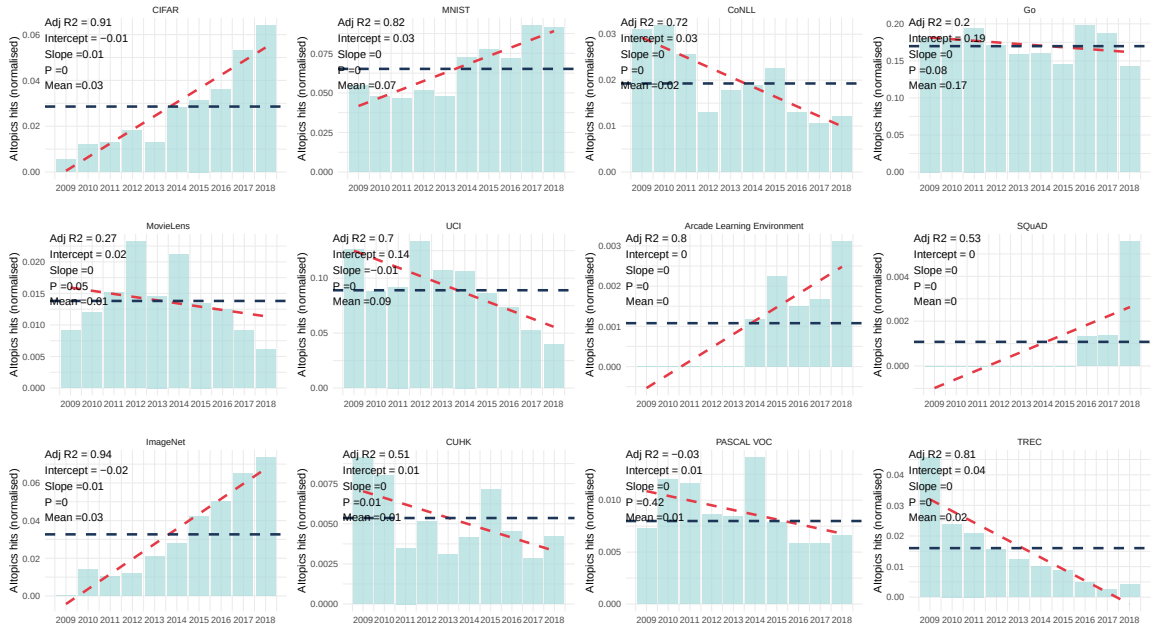



Figure 4: Average rate of activity level or intensity (dark blue dashed line) and linear fit (red dashed line) for a set of illustrative AI benchmarks over the period 2008-2018.

## 6. Results

Before presenting the results of the AI exposure score, we illustrate the process of the development of the framework through intermediate results of the mapping of abilities to tasks and the mapping of AI benchmarks to abilities. More detailed results of the annotation exercise for the assignment of abilities to tasks are shown in Appendix B.

### 6.1 Tasks and Cognitive Abilities

The ability-specific intensity matrix $\mathbf{W}(119 \times 14)$ shows for every occupation the relevance of each cognitive ability against the remaining cognitive abilities. For instance, although it does not show whether the ability *communication* (CO) is in absolute terms more required in one occupation over another, it can tell us whether *communication* (CO) is more relevant relative to other cognitive abilities in one occupation over the other occupations. Thus, it informs about which occupation can be more affected by increased research intensity in *communication* (CO) or any other of the 14 cognitive abilities.

Note that we take the part of the framework that specifies work contents in terms of task from the task framework presented in Fernandez-Macias and Bisello (2020), according to which tasks are, at the highest level of abstraction, classified according to the type of object upon which they can operate: 1) social relations, 2) ideas or information, and 3) physical objects (see Section 3.1). The cognitive abilities approach comes from a very different perspective, but can be very complementary using the mapping that we present in this paper. This is mostly because any kind of human activity (including work and tasks) requires the use of some cognitive ability. There-

fore, we translate the high level categorisation of work tasks to cognitive abilities by sorting each ability according to the objects that they operate on into one of the following three categories: (1) dealing with **people**: emotion and self-control (**EC**), mind modelling and social interaction (**MS**), metacognition and confidence assessment (**MC**), mind modelling and social interaction (**MS**); (2) dealing with **ideas** or information: comprehension and expression (**CE**), planning, sequential decision-making and acting (**PA**), memory and processes (**MP**), attention and search (**AS**), conceptualisation, learning and abstraction (**CL**), quantitative and logical reasoning (**QL**); and (3) dealing with (physical or virtual) objects or **things**: sensorimotor interaction (**SI**), navigation (**NV**), visual processing (**VP**), auditory processing (**AP**). In the following we abbreviate these categories to (1) *people*, (2) *ideas*, and (3) *things*.

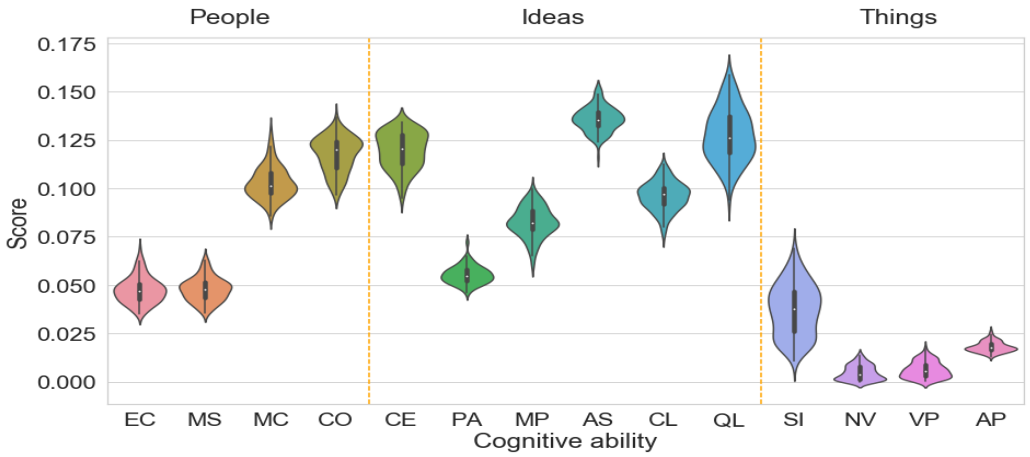



Figure 5: Distribution of ability-specific scores across occupations.

Figure 5[24], which plots the elements in matrix $\mathbf{W}(119 \times 14)$, reveals two aspects about the relevance of cognitive abilities across occupations: variation and level. In terms of variations, the length of the bands reveal which abilities *vary* most in terms of their relevance across occupations. Many abilities have similar relevance variation where *quantitative reasoning* (QL), and *sensorimotor interaction* (SI) depict clear exceptions. Moreover, the variation in relevance for *communication* (CO) as well as *comprehension* (CE) is also noteworthy. This means that, from the perspective of work content, high AI research intensity in QL, SI, CO or CE exhibits the largest differences in terms of likelihood of AI impact across occupations, while AI research intensity in the other abilities could potentially affect most occupations equally. Hence, the higher variations in the relevance of *people* or *ideas* abilities imply that any AI that performs well on these types of cognitive abilities yields higher variation in occupational exposure.

In terms of relevance levels, the figure shows that for most occupations, abilities of the categories *people* and *ideas* are more relevant at the workplace than *things* abilities.[25] More specifically,

24. For a more detailed view on the distribution of these ability-specific scores within occupations, we present in Figure 9 - Appendix F the same scores of matrix $\mathbf{W}(119 \times 14)$ for nine selected ISCO-3 occupations.
25. There are some reasons (related to work contents and the mapping from cognitive abilities to tasks) to assume why the present framework does not capture well the occupational relevance of the *things* abilities NV, VP, SI. First,

the most relevant abilities are *quantitative reasoning* (QL), *comprehension* (CE), *communication* (CO), *attention and search* (AS) as well as *metacognition* (MC). Furthermore, we find that *conceptualisation* (CL) and *memory processing* (MP) present high relevance levels for most occupations. In contrast, of the abilities that deal with *things* only *sensorimotor interaction* (SI) has high relevance levels for a large share of occupations, whereas other abilities of that category do not appear relevant except for a small minority of occupations. Thus, the higher occupational relevance of *ideas* and *people* abilities compared to *things* abilities implies that any AI that performs well on *ideas* or *people* abilities yields more occupational exposure than an AI that performs well on *things* abilities.

### 6.2 AI Research Intensity in Cognitive Abilities

Vector $\mathbf{a}(14 \times 1)$ indicates for each cognitive ability the relative AI research intensity. We illustrate this vector in Figure 6. We see that most AI research activity can be attributed to *visual processing* (VP), *attention and search* (AS), *comprehension, compositional expression* (CE), *conceptualisation, learning and abstraction* (CL) and *quantitative and logical reasoning* (QL).

Figure 6 shows the computed AI research intensity for each cognitive ability for every two-year period from 2008 to 2018. The figure shows that AI is currently having a larger relative intensity on those cognitive abilities that rely on memorisation, perception, planning and search, understanding, learning and problem solving, and even communication; smaller influence on those more ambient-related abilities belonging to the *things* category introduced above, namely, navigation and interaction with the environment. Since “intensity” depends on the level of activity on *AI topics*, this would mean that there is a lower amount of documents related to those benchmarks dealing with (physical or virtual) objects or things, but also, although to a lesser extent, due to a more limited number of robotics benchmarks, which are usually more difficult to build and maintain. Note that the focus of this paper is AI, which includes some areas in robotics (such as *cognitive robotics*) but not others.

note that the framework from which we draw information on work contents (see Table 1) was only later updated to contain the indices "navigation" and "processing of uncodified information". Considering these updates to the task framework would potentially yield more annotations (and consequently higher relevance scores) for NV, VP and SI. However, AI research intensity is comparatively low for navigation (see Figure 6). So higher occupational relevance scores for NV would not have had a strong impact on the final AI exposure score. Secondly, when mapping abilities to tasks, abilities were only assigned to tasks when absolutely necessary. In many cases, assigning VP or AP was circumvented by assigning SI.

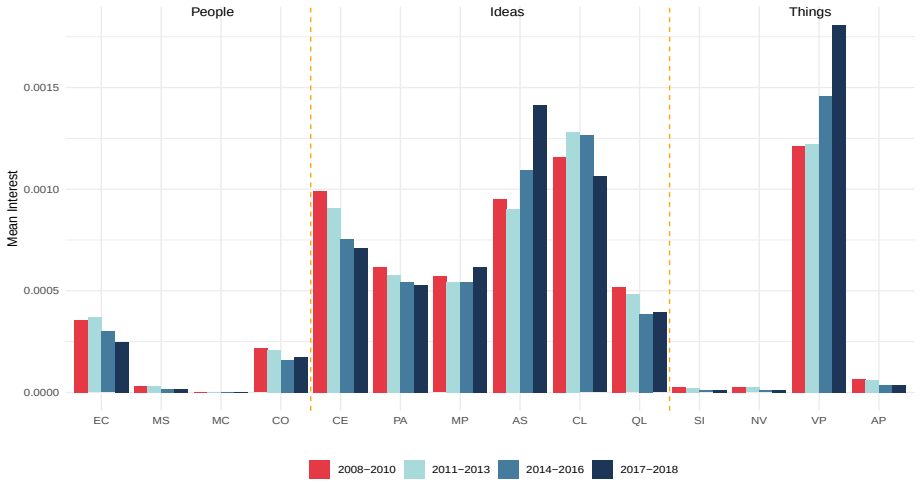



Figure 6: Relevance per cognitive ability weighted by (average) rate intensity for different periods of years over the last decade (2008-2018).

We also see almost no research intensity on those abilities related to the development of *social interaction* (MS) and *metacognition* (MC). This may be due to the lack of suitable benchmarks to evaluate the interactions of agents (human and virtual) in social contexts; as well as the challenge (today) of developing agents able to properly perform in social contexts with other agents having beliefs, desires and intentions, coordination, leadership, etc. as well as being aware of their own capacities and limits.

Note that Figure 6 also shows trends over the years for each cognitive ability. There is a clear increasing trend in *visual processing* (VP) and *attention and search* (AS), while other abilities remain more or less constant (MP, SI, AP, CO, CL and MS) or have a small progressive decline (PA, CE, EC and QL). Note that these values are relative. For instance, PA, CE or QL have decreased in proportion to the rest. In absolute numbers, with an investment in AI research that is doubling every 1-2 years (Shoham et al., 2018), all of them are actually growing. Thus the figure shows that imbalances are becoming more extreme.

### 6.3 AI Exposure Score

This section describes the results from the combination of all three layers of the framework: (1) tasks, (2) cognitive abilities, and (3) AI benchmarks in terms of occupations (see Section 4.3 for the corresponding methodology). Using the AI research intensity scores from 2018, we compute matrix $\mathbf{V}(119 \times 14)$, the ability-specific matrix of AI exposure scores. As mentioned in Section 4.3 this score indicates which of the studied occupations are relatively more likely to be affected by AI research intensity through which cognitive ability. Again, we focus on the nine selected occupations specified above (but slightly abbreviate the occupational titles for readability).

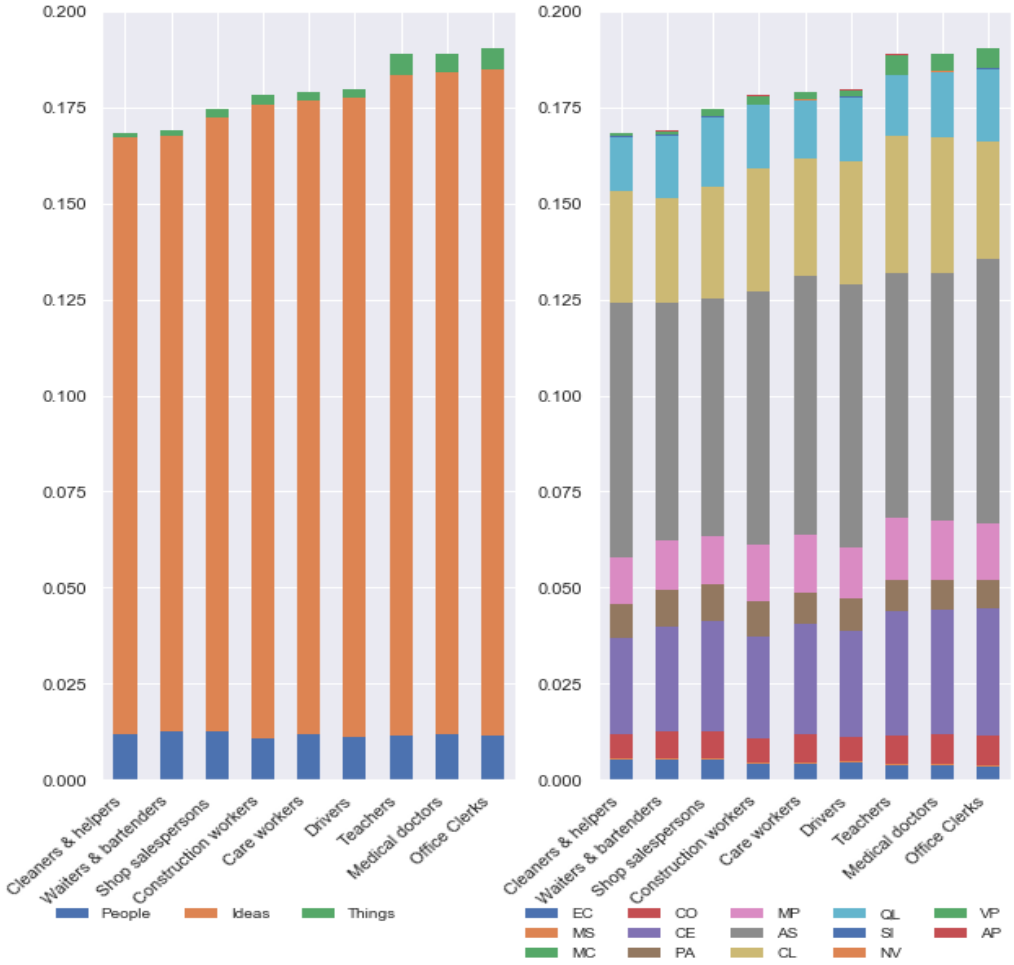



Figure 7: Ability-specific AI exposure scores for selected occupations. Left: grouped by people, ideas or things abilities. Right: detailed for the 14 abilities.

Figure 7 depicts the computed AI exposure score differentiated by cognitive ability categories on the left and by cognitive abilities on the right for nine selected ISCO-3 occupations (see also matrix $\mathbf{V}(119 \times 14)$): general office clerks, shop salespersons, cleaners and helpers, medical doctors, personal care workers in health services, primary school and early childhood teachers, heavy truck and bus drivers, waiters and bartenders, building and related trades in construction. We sort the occupations left to right from lowest to highest AI exposure score. First, the figure shows that general office clerks, medical doctors and teachers are more exposed to AI research intensity than occupations that require comparatively lower skills such as cleaners, waiters or shop salespersons. Second, Figure 7 shows that most of AI exposure is driven by its impact on tasks that require abilities that deal with *ideas*, such as *comprehension* (CE), *attention and search* (AS) as well as *conceptualisation* (CL). This is not because we assign more cognitive abilities (6) to the *ideas* category than to the other categories (each 4), since the smallest exposure score from the *ideas* abilities (in most cases *planning and action* (PA)) is still often higher than the highest exposure score from the *people* category (*communication*). Compared to this, the exposure

scores in the *things* category are negligibly small. That is, not much AI exposure can be expected through basic processing abilities, such as *visual processing* (VP) or *auditory processing* (AP), nor through *mind modelling and social interaction* (MS). Although some exposure also occurs through AI research intensity in *communication* (CO). However, our findings based on the tasks and occupation data indicate a relatively high need for *people* abilities in most occupations and a relatively low need for abilities dealing with *things*. Equivalently, the findings on AI research intensity suggest high activity in AI areas that contribute to abilities dealing with *things* but also to the abilities with the highest exposure score mentioned above, and low activity for abilities dealing with *people*.

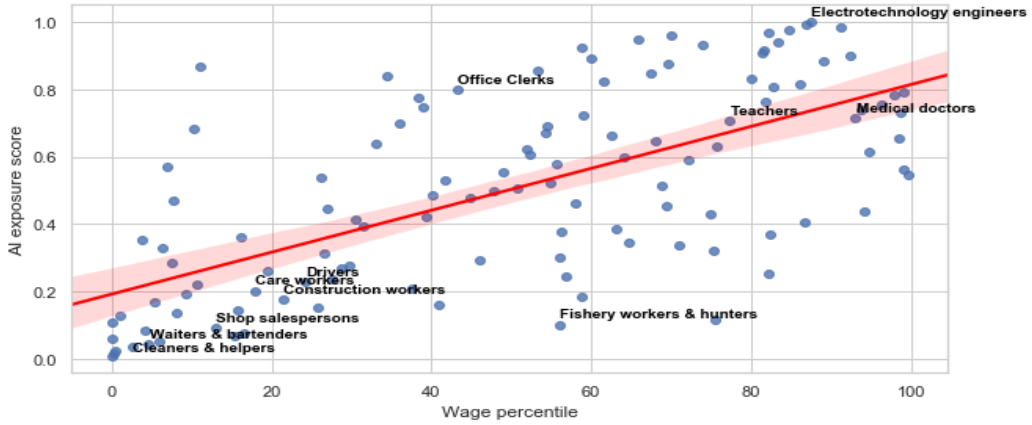



Figure 8: Scatterplot and best fit line, AI exposure score (percentiles) against wage percentiles
Source: Structure of Earnings Survey 2014

Finally, to compute a single AI exposure score for each occupation, we take the sum over the columns of matrix $\mathbf{V}(119 \times 14)$ and take the percentiles of this sum. This score is presented in Table 8 of Appendix E. Note that this score does not represent a percentage but it can be used to infer a ranking between occupations in terms of AI exposure. To show how some uncertainty in the annotation of abilities to tasks might translate in the computation of the final AI exposure score, the table also presents values based on the upwards and downwards perturbation of the values of matrix $\Omega$. As already seen in Figure 7, the table suggests higher impact for occupations, such as medical doctors, school teachers, or electrotechnology engineers. This may be surprising, because these are occupations that were less affected by previous waves of automation. However, we would like to emphasise that our analysis does not focus on the automation potential of AI, but on what kinds of task content and occupations are more likely to be affected by current developments of AI. Although we do not explicitly discuss this in this paper, it is likely that the effect we are talking about is one of labour complementarity rather than substitution: after all, we are speaking about applications of AI that expand human cognitive abilities in key areas such as comprehension, attention and search, and conceptualisation.

Furthermore, in Figure 8 we plot the AI exposure score (in percentiles) against average wage percentiles of each studied occupation.[26] The figure shows a positive relationship between wages and AI exposure. That is, high-income occupations seem more likely to be affected by AI research intensity, than low-income occupations. It is well possible that some very basic skills that are taken for granted for every human, such as naive physics (moving around and manipulating objects), language fundamentals (using language at the basic level, following orders) and naive psychology (understanding agency in other people), which are usually captured under the term common sense in AI (Davis and Marcus, 2015), are not fully represented in the descriptions that are used by the work intensities. But these may be required by all occupations and cover things, ideas and people. We further discuss the implications of these results in Section 7.

### 6.4 Validation

Despite the wide agreement that the latest advances in AI will have disruptive repercussions on the labour market, there is very little evidence for actual integration of AI in the labour market. Brynjolfsson et al. (2018) call this the modern productivity paradox: We see AI achieving superhuman abilities in some tasks, and companies that heavily invest in AI among the highest valued in the world[27] but we do not measure it in relevant productivity statistics. In fact, we are currently measuring the impact of AI on labour markets before the necessary investments to enable the implementation and diffusion of AI in the labour market may have occurred (Brynjolfsson et al., 2021). Therefore, this framework can be considered as a glimpse of how the occupational impact of AI may potentially look like, although based on data-supported assumptions on AI progress and work content. Thus, we have to resort to alternative measures providing evidence to validate the present framework.

A potential alternative measure for the occupational impact of AI may be the creation rate of new job titles in occupations that are most affected by AI. Indeed, based on US data, Acemoglu and Restrepo (2018) find that around 60 percent of newly created jobs between 1980 and 2015 are associated with faster employment growth in occupations with new job titles. However, there are many other reasons than the creation of new tasks (and consequently new or restructured work content) that drive changes in job titles, such as regulations, cultural change and social convention (Fernández-Macías et al., 2018).

Other measures of the occupational impact of AI may relate to investments in research and development or the count of AI related patents since over the past decades such measures have been used by the economic literature as proxy for the level of innovation (Coad and Rao, 2011; Bogliacino et al., 2012). Although there is significant evidence that innovative firms hire more workers, there is not yet evidence on how the investment of firms on innovation activities affects the level of total employment. At firm level, it exists a clear distinction between the effects of product innovation and process innovation, the former being much stronger than the latter, particularly for high-tech industries (Calvino and Virgillito, 2018). The potential disruption of AI in the organisation of work evidently rely upon process innovation. Making a parallel between

26. We obtain data on wage percentiles of 12 EU member states at ISCO-3 level from the structure of earnings survey 2014: `https://ec.europa.eu/eurostat/web/microdata/structure-of-earnings-survey`

27. See `https://www.statista.com/topics/4213/google-apple-facebook-amazon-and-microsoft-gafam/`, last access: December 6, 2020.

the process innovation introduced by AI and industrial robots, for the latter Acemoglu and Restrepo (2018) look at the effect on US local labour markets; if anything they find a reduction of employment, suggesting that while process innovation could be beneficial at firm level, its effect is less clear-cut for the general labour market. Similarly, such ambiguity remains when using patenting activity to measure the effect of product innovation from the firm to the labour market (Ciarli et al., 2018). Additionally, the use of raw patent counts, although broadly used in the economic literature, is subject to longstanding debate about its shortcoming and several sources of bias (Archibugi, 1992; Dosi, 1988; Griliches, 1998). As for example, different sectors and countries have different patenting behaviours as well as big companies and small firms. Furthermore, the simple raw count of patents does not give any indications about their quality, as patenting processes do not involve any quality assessment. Patent citations have been used in order to correct for the lack of implicit quality judgement in raw counts of patents. However, the use of patent citations as an indicator also presents several limitations (Jaffe and de Rassenfosse, 2017).

In order to validate the information gathered from AI topics as a measure of research intensity, as opposed to the use of patents or investment in research and development, we have performed the same quantitative analysis using data obtained from Google Trends[28]. Google Trends provides a normalised measure of search volume (i.e., popularity) for a given search term over a selected period of time. Google Trends has more technical constraints compared to AI Topics: all searches are scaled to the highest volume topic in your query and there is a limit of five topics per query. This can be solved by using a control topic in each search (two different queries are comparable if they have the same largest topic). Despite these issues, Google Trends is easily accessible, freely available, and broadly used in the literature to analyse interest in a keyword or topic over time (see e.g., Preis et al. (2013); Nuti et al. (2014)). We extract the aggregated trends per benchmark in the same way and for the same period as we did for AI topics. We calculate Spearman's correlation between both, giving a result of 0.43. This implies a moderate correlation between the results from AI topics and Google Trends. However, it should be noted that more than 80% of the benchmarks got a relative search interest equal or close to 0, representing a very low significance. If we limit our correlation analysis to those benchmarks with non-zero interest according to Google Trends, the correlation increases to 0.68, a stronger correlation, supporting the methodology. The above results suggest that Google Trends may be considered as an alternative to AI topics, although AI topics is a more comprehensive source of data.

For a qualitative validation of the computed AI exposure score, we compare this score to three other occupational AI exposure scores in the literature: (1) Brynjolfsson et al. (2018) (BMR)[29], (2) Webb (2020)[30], and (3) Felten et al. (2018) (FRS)[31]. Note that the main difference between these approaches and the one taken in this paper is their reliance mainly on US task data (O*NET) and their direct link between a measure of AI capabilities and task descriptions. Besides, each framework uses a different source for measuring AI capabilities. While Brynjolfsson et al. (2018) rely on a rubric that determines the suitability of a task for machine learning, Webb (2020) assesses overlap between AI-related task and patent descriptions. The approach taken by Felten

28. https://trends.google.com/
29. Source of material for scores: https://www.aeaweb.org/articles?id=10.1257/pandp.20181019
30. Source of material for scores https://www.michaelwebb.co/
31. Source of material for scores https://www.aeaweb.org/articles?id=10.1257/pandp.20181021

et al. (2018) is closest to ours as their measure derives from an AI benchmarking platform (see Section 2 for a discussion of the differences in the approaches) and they link these benchmarks to work-related capabilities. Since all three scores are encoded in O*NET's occupational classification (SOC), we first apply a crosswalk[32] to obtain the scores at the 4-digit ISCO level. Next, we average each score to 3-digit ISCO level and transform the scores to percentiles.

We finally analyse the Spearman rank correlations between the percentiles of the AI exposure score in this study and the percentiles of the other AI exposure scores. We find that each score coming from the three studies is significantly correlated (with p-values $\leq 0.001$) with the AI exposure score in our study. Although the correlations are at relatively low levels with $\rho$ values below 0.5 in all cases, some occupations have relatively similar values in all approaches. Moreover, our score has the highest correlation with the FRS score, which is based on an approach closest to the one taken in this paper (scatter plots and more detailed analysis around Figure 10 in Appendix F).

## 7. Discussion

The societal perception of innovation and technological progress has always been ambiguous (Mokyr et al., 2015). While there is wide agreement that AI will change the way we work and the structure of our labour markets, there is a lot of uncertainty about the direction that these changes will take (Shoham et al., 2018). Those who perceive such disruption in a negative light project a future where human labour will be mostly replaced by robots, and technology will shape the rules of society in a dystopia of technological determinism. Others have a more positive perception according to which technological progress, as it has done in the past, will mostly enhance human labour and create new and better jobs (Cockburn et al., 2018). It has become clear that there are many effects to consider that determine the outcome of these economic (and consequently societal) processes (Brynjolfsson and Mitchell, 2017).

The present analysis is limited to the technical potential of AI (i.e., the things that AI could potentially do at work). We can use this approach to highlight occupations and abilities involved where AI could play a role. However, there are other factors that affect AI diffusion that are not covered by our framework. For instance, some complementary conditions and restructuring of business processes may be necessary to enable the integration of AI in the workplace (Brynjolfsson et al., 2018; Brynjolfsson and Mitchell, 2017). Another important factor is the relative cost of labour vs. AI. For instance, replacing labour by AI would only be economically feasible if the effective costs of performing the task with machines are lower (Acemoglu and Restrepo, 2018). This relates to the elasticity of substitution between AI and labour, and correspondingly the role of aggregate demand in the diffusion of AI. Productivity enhancing technologies (such as AI) could result in an increase in employment if price effects cause increased demand (Bessen, 2019). However, AI does not guarantee demand growth. Gries and Naudé (2020) show this by incorporating the task-based approach into an endogenous growth model. In fact, a high elasticity of substitution between AI and labour decreases the GDP share of labour income (and increases the share of AI providers) and consequently reduces aggregate demand causing a deceleration in AI diffusion. Thus, it is not given that technological feasibility of AI will automatically cause AI diffusion; our

32. Source: `https://ibs.org.pl/app/uploads/2016/04/onetsoc_to_isco_cws_ibs_en1.pdf`

results have to be interpreted in light of this limitation. Nevertheless, based on the assumption that the present AI exposure score and AI diffusion are correlated at the occupational level, we can use the results of this study to shed light on some aspects of the relationship between AI research intensity and labour markets.

These findings suggest that AI (as an emerging technology) will probably not have the type of labour market polarisation effects that some people associate with the recent wave of computerisation. According to some studies, previous waves of technological progress led to a polarisation on the labour market where the automation of middle-skill occupations pushed middle-skill workers to either low- or high-skill occupations (Autor et al., 2003; Goos et al., 2014). In this line of research, the classification into high-/middle-/low-skill occupations is done on the basis of wage percentiles or educational levels, where this puts medical doctors and engineers into the high-, sales and office clerks into the middle-, and drivers or cleaners into the low-skill occupations. In contrast, our findings (according to Figure 8) suggest relatively high AI exposure for high-skill occupations but relatively low AI exposure for low-skill occupations such as drivers or cleaners, while there seems to be no clear pattern for middle-skill occupations (e.g. high exposure for general office clerks but low exposure for fishery workers and hunters). This is in line with findings by Brynjolfsson et al. (2018) and Webb (2020) (see also Figure 10).

This can have different implications for occupational change (and consequently inequality) depending on whether AI exposure is labour replacing, or labour enhancing. If this effect is in fact a labour-replacement one, it could potentially lead to unpolarising effects and a reduction in income inequality (Webb, 2020). If this effect is a labour-enhancing one, it could imply a significant expansion of productivity for high-skilled occupations, potentially leading to occupational upgrading effects and an expansion of income inequality (very much like the traditional hypothesis of skills-biased technological change (Acemoglu, 2002)).

Furthermore, because we break down the effect of AI research intensity into 14 different abilities, our findings show that AI progress could affect how specific skills are rewarded (e.g., in terms of wages and working conditions) on the labour market.[33] The finding of low exposure through *people* abilities versus high exposure through *ideas* abilities is parallel to Deming (2017), who explores the relationship in the labour market returns to social skills and, what he calls, *cognitive skills* to which here we refer to as *analytic skills*.[34] In a way, Deming (2017)'s social and analytic skills are equivalent to the *people* and *things* abilities in the present paper. More specifically, we use social skills to interact with people and we use abilities that deal with ideas (such as *conceptualisation* (CL), *quantitative reasoning* (QL) or *comprehension* (CE)) in areas that require analytic skills. Deming (2017) finds that social and analytic skills are complements rather than substitutes. That is, an increased labour demand for analytic skills, which increases wages for people with analytic skills, leads to an increased labour demand for people that, in addition to analytic skills, also have strong social skills. In addition, we find that many labour market tasks require high levels of *people* as well as *ideas* abilities but AI exposure occurs mostly through

33. Note that we define cognitive abilities and skills as two distinct properties. However, since cognitive abilities explain a part of skills we focus here on comparing the parallels of both properties.
34. In (Deming, 2017) *cognitive skills* are skills in the areas of maths, statistics, engineering and science. To avoid this being confused with this study's term "cognitive abilities", we change the name to analytic skills.

*ideas* abilities only. That is, if the higher AI exposure eventually causes increases in the efficiency of tasks that require *ideas* abilities, prices for the products of these tasks may decrease, causing increased demand for these products and consequentially increased labour demand for these tasks (Bessen, 2019). If these tasks also contain a high need for *people* abilities, of which we find that they are not likely to be affected by AI in the near future, we can expect an increase in the wages for workers that combine their strong *ideas* abilities with strong *people* abilities.

A combination of increased wages for some occupations and tasks can lead to some AI technologies being more valuable, and more investment in research and computation being justified for them. For instance, the recent progress in massive language models in AI (see e.g., Brown et al. (2020); Hendrycks et al. (2020)), which relies on expensive computation, is at the core strategy of major AI laboratories and their alliance with tech giants (e.g., OpenAI with Microsoft). They are producing more effective presentations as cognitive services that may end up having new effect on abilities such as CO and CE, if they increase the productivity of these high-wages occupations related to ideas and people.

## 8. Conclusion

In this paper we developed a framework that allows for the analysis of the impact of artificial intelligence on the labour market. The framework combines occupations and tasks from the labour market with AI research intensity through an intermediate layer of cognitive abilities. This approach allows to accurately assess the technological potential of AI in work-related tasks and corresponding occupations. We use the framework to rank selected occupations by potential AI impact and to show the abilities that are most likely exposed to AI progress. We find that some jobs that were traditionally less affected by previous waves of automation may now be subject to relatively higher AI exposure. Moreover, we find that most of the AI exposure occurs through abilities that we use to deal with ideas. In light of the digital transformation and the rise of AI, these findings can help policymakers in directing their response in the form of education and (re-)training policies, and inform individuals in their career choice. In addition, breaking down the occupational effect to tasks and cognitive abilities can inform employers in the restructuring of occupations and tasks as the framework also informs about the particular capacities (abilities) within a task that may be supported by AI.

The focus on abilities, rather than task characteristics, goes beyond measuring the substitution effect of AI. Most AI applications are built to perform certain abilities, rather than execute full work-related tasks and most tasks will require multiple abilities to be executed. Identifying the specific abilities that can be performed by AI gives a broader understanding of the impact of AI. Relying on AI field benchmarks that are used as orientation by AI researchers and other AI industry players makes the framework adoptable to future developments in AI research. As mentioned above, AI exposure does not necessarily mean automation. So, our findings do not imply that all tasks that mostly require abilities to deal with ideas will be automated, as AI exposure can also mean that the way a task is being performed is just restructured.

Moreover, we find that most occupations map to abilities that deal with people very relevantly, while AI progress has a stronger effect on abilities that deal with ideas. Corresponding labour

market processes could potentially increase the demand for workers with strong *people* abilities. Overall, we can be much more certain about the capacity of AI to transform jobs than about its capacity to destroy them.

This framework can also be used for counterfactual simulations of changes in AI research activity. For instance, the framework can be used to uncover the AI benchmarks that contribute to people-related abilities in order to simulate an AI exposure score with more research activity in benchmarks that contribute to *people* abilities. In addition, this framework is useful if in the future new occupations are created that require different/new tasks and consequently require different/new cognitive ability profiles. In this case the framework can be used to reveal the benchmarks for which more research activity is needed to address these changing requirements.

There are many aspects about the future that escape this work, as happens with many other studies about AI and the future of work. For instance, while the results in scaling laws (Henighan et al., 2020) seem to be pushing that progress can continue with more massive deep learning architectures in several domains using more computation and resources, there are opposing factors such as their impact on the sustainability of AI, the public opinion about automation and an increasing sense of distrust in AI, which may affect some professions more than others, and fewer jobs may really be exposed if that societal and regulatory effect takes place in the future.

In future work, other task characteristics such as work organisation could be integrated into the framework. This will allow us to measure and distinguish the impact of AI through newly acquired technical capabilities and the automation potential of tasks. Moreover, the measurement can be refined as more data on the relevance of specific work-related tasks as well as new benchmarks on AI progress arise. Overall, this framework can help bridge the gap between research in labour and AI.

## Appendix A. Cognitive Abilities Rubric

As described in the main text, the following cognitive abilities are integrated from different sources in psychology, animal cognition and artificial intelligence. We include the description of the ability and a rubric to help map task to abilities.

**MP: Memory processes**: 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.

**SI: Sensorimotor interaction**: 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.

**VP: Visual processing**: 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.

**AP: Auditory processing**: 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.

**AS: Attention and search**: 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.

**PA: Planning, sequential decision-making and acting**: 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.

**CE: Comprehension and expression**: 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..

**CO: Communication**: 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.

**EC: Emotion and self-control**: 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.

**NV: Navigation**: 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..

**CL: Conceptualisation, learning and abstraction**: 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.

**QL: Quantitative and logical reasoning**: 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.

**MS: Mind modelling and social interaction**: 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.

**MC: Metacognition and confidence assessment**: 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.

## Appendix B. Mapping Abilities to Tasks

In this section we summarise the results of the annotation of abilities to tasks. The annotations of each round are put together in a Matrix $\Omega(59 \times 14)$ for 59 tasks ($t \in \mathrm{T}$) and 14 cognitive abilities ($a \in \mathrm{A}$), where each cell in $\Omega$ ($\omega_{t,a}$) represents the sum over all annotations of a respective round. On the task level we describe our results by number of assigned abilities and consensus. An ability is assigned to a task if at least two annotators assigned a 1 for a respective task-ability cell. That is, for each task $t$, we define the number of assigned abilities as:

$$\mathrm{S}(t) = \sum_{a} \left[\omega_{t,a} \geq 2\right]$$

where [P] are the *Iverson brackets* : [P] is defined to be 1 if P is true, and 0 if it is false.

We also compute the level of consensus among respondents using a geometry-based disagreement measure following on from the work of Saari (2008); Claveria et al. (2019). Here, the authors define a framework to proxy economic uncertainty or to determine the likelihood of discrepancy among respondents. In our setting, we assume a dichotomous questionnaire with N = 2 reply options (e.g., ability is assigned or not to a task), and $\mathrm{R}_{i,a}$ denoting the aggregate percentage of responses in category $i \in \{1,0\}$ for a specific ability $a \in \mathrm{A}$. As the sum of R adds to 100, a natural representation of the vector containing all the information from the respondents for a given ability $a$ is as a point on a 1-dimensional (2 vertexes) simplex (Coxeter, 1961). Note that, while each of the N vertexes corresponds to a point of maximum consensus, if the point is near the barycenter, there would be a maximum discrepancy among the respondents. We can then compute the consensus between respondents as the relative weight of the distance of each point to the barycenter, formalised as:

$$\mathrm{C}_a = \sqrt{\frac{\sum_{i=1}^{\mathrm{N}} (\mathrm{R}_{i,a} - \frac{100}{\mathrm{N}})^2}{\frac{(\mathrm{N}-1)}{\mathrm{N}}}}$$

As can be seen in Table 5, we find that the annotators become stricter with their assignments of cognitive abilities to tasks in the second round. In addition, consensus in assignments increases from on average 80.65% to 87.6% from one round to the next.

| | S | | | C | | |
|---|---|---|---|---|---|---|
| | round 1 | round 2 | diff. | round 1 | round 2 | diff. |
| **Average** | 6.03 | 5.34 | -0.69 | 80.65% | 87.7% | 7.05 p.p |
| **Min** | 0 | 0 | 0 | 57.14% | 69.05% | 11.91 p.p |
| **Max** | 13 | 10 | -3 | 100.00% | 100.00% | 0 |

Table 5: Difference in annotations between round 1 and round 2

## Appendix C. List of Tasks

| | |
|---|---|
| 1 | Task involving tiring or painful positions |
| 2 | Lifting or moving people |
| 3 | Carrying or moving heavy loads |
| 4 | Standing |
| 5 | Static Strength |
| 6 | Dynamic Strength |
| 7 | Trunk Strength |
| 8 | Arm-Hand Steadiness |
| 9 | Manual Dexterity |
| 10 | Finger Dexterity |
| 11 | Oral Comprehension |
| 12 | Written Comprehension |
| 13 | Oral Expression |
| 14 | Written Expression |
| 15 | Read letters, memos or e-mails |
| 16 | Read bills, invoices, bank statements or other financial statements |
| 17 | Write letters, memos or e-mails |
| 18 | Read directions or instructions in your job |
| 19 | Read manuals or reference materials? |
| 20 | Read diagrams, maps or schematic in your job |
| 21 | Have to write reports |
| 22 | Have to fill in forms |
| 23 | Read articles in newspapers, magazines or newsletters |
| 24 | Read articles in professional journals or scholarly publications |
| 25 | Read books |
| 26 | Write articles for newspapers, magazines or newsletters |
| 27 | Mathematical Reasoning |
| 28 | Number Facility |
| 29 | Calculate prices, costs or budgets |
| 30 | Use or calculate fractions, decimals or percentages |
| 31 | Use a calculator either hand-held or computer based |
| 32 | Prepare charts, graphs or tables |
| 33 | Use simple algebra or formulas |
| 34 | Use more advanced math or statistics |
| 35 | Learning new things |
| 36 | Deductive Reasoning |
| 37 | Inductive Reasoning |
| 38 | Information Ordering |
| 39 | Solving unforeseen problems on your own |
| 40 | Apply your own ideas in your work |
| 41 | Originality |
| 42 | Performing for or Working Directly with the Public |
| 43 | Selling a product or selling a service |
| 44 | Advising people |
| 45 | Persuading or influencing people |
| 46 | Negotiating with people either inside or outside your firm or organisation |
| 47 | Persuasion |
| 48 | Negotiation |
| 49 | Selling or Influencing Others |
| 50 | Resolving Conflicts and Negotiating with Others |
| 51 | Instructing, training or teaching people |
| 52 | Making speeches or giving presentations in front of five or more people |
| 53 | Instructing |
| 54 | Training and Teaching Others |
| 55 | Coaching and Developing Others |
| 56 | Manage or supervise other employees |
| 57 | Planning the activities of others |
| 58 | Coordinating the Work and Activities of Others |
| 59 | Guiding, Directing, and Motivating Subordinates |

Table 6: Lists of Tasks used in Mapping

## Appendix D. List of AI Benchmarks

| Benchmark | Mean intensity | Benchmark | Mean intensity | Benchmark | Mean intensity | Benchmark | Mean intensity |
|---|---|---|---|---|---|---|---|
| 20NEWS | 0.00498666 | Event2Mind | 0.000004 | MR | 0.042382 | Shogi | 0.00029975 |
| 300W | 0.00064231 | Fashion-MNIST | 0.001135 | MRR | 0.004226 | SighanNER | 0.00000000 |
| ACE 2004 | 0.00011392 | FB15k | 0.000759 | MS COCO | 0.001450 | SimpleQuestions | 0.00154478 |
| ACE 2005 | 0.00063344 | FB15k-237 | 0.000153 | MS MARCO | 0.000415 | Sintel | 0.00018307 |
| ADE20K | 0.00012735 | FCE | 0.000201 | MSRA | 0.002307 | SK-LARGE | 0.00000405 |
| Aerial-to-Map | 0.00000000 | FDDB | 0.000052 | Multi-Domain Sentiment Dataset | 0.000759 | SLAM 2018 | 0.00000809 |
| AEROBCOMP | 0.00000000 | FFHQ | 0.000004 | MultiMNIST | 0.000124 | SNLI | 0.00047133 |
| AFAD | 0.00000000 | FGNET | 0.000557 | MultiNLI | 0.000150 | Sogou News | 0.00005982 |
| AFLW | 0.00011200 | FGVC Aircraft | 0.000057 | MultiRC | 0.000008 | spider | 0.00275807 |
| AG News | 0.00017801 | fisher WER | 0.000000 | Mushroom | 0.007158 | SQuAD | 0.00075022 |
| AI2 Kaggle Dataset | 0.00000000 | FLIC | 0.000180 | Music domain | 0.000850 | SR11Deep | 0.00000000 |
| Amazon Review | 0.00094449 | Flixster | 0.000882 | MUV | 0.000432 | SST | 0.00261894 |
| ANGRY-BIRDS | 0.00019157 | Florence | 0.003080 | NABirds | 0.000020 | Stanford Cars | 0.00006772 |
| Annotated Faces in the Wild | 0.00002672 | Flowers-102 | 0.000252 | NarrativeQA | 0.000058 | Stanford Dogs | 0.00033553 |
| Arcade Learning Environment | 0.00088491 | GENIA | 0.001328 | NELL | 0.002441 | STARE | 0.00037779 |
| bAbi | 0.00004494 | GigaWord | 0.000357 | NER | 0.008085 | Static Facial Expressions in the Wild | 0.00000000 |
| Bing News | 0.00009736 | GLUE | 0.003006 | Netflix | 0.020838 | STL-10 | 0.00191744 |
| BIWI | 0.00008235 | Go | 0.172822 | New York Times Corpus | 0.000721 | Story Cloze Test | 0.00004344 |
| BlogCatalog | 0.00084899 | Google Dataset | 0.001131 | NewsQA | 0.000139 | STS | 0.00359873 |
| Bosch Small Traffic Lights | 0.00000405 | Google Street Images | 0.000034 | North American English | 0.000014 | SUBJ | 0.00524186 |
| BotPrize | 0.00021779 | GTA V | 0.000045 | Noun Phrase Canonicalization | 0.000000 | SUN-RGBD | 0.00009389 |
| BP4D | 0.00005449 | GTSRB | 0.000375 | NYU Depth v2 | 0.000311 | SVHN | 0.00392847 |
| BPI challenge | 0.00004494 | GVGAI | 0.000047 | NYU Hands | 0.000000 | SVNH-to-MNIST | 0.00000000 |
| BRATS | 0.00013988 | HANDS 2017 | 0.000000 | Occluded LINEMOD | 0.000000 | SWAG | 0.00010718 |
| BSD* | 0.00267574 | Helpdesk | 0.000621 | OCCLUSION | 0.015062 | Switchboard | 0.00214692 |
| BUCC | 0.00003076 | HIV dataset | 0.000448 | Ohsumed | 0.003953 | SYNTHIA | 0.00011521 |
| BUS 2017 | 0.00000000 | HotpotQA | 0.000000 | OMNIGLOT | 0.001333 | T-LESS | 0.00014327 |
| CACD | 0.00002522 | Human3.6M | 0.000198 | One Billion Word | 0.000602 | TACRED | 0.00001214 |
| CACDVS | 0.00001713 | Hutter Prize | 0.000429 | OntoNotes | 0.000543 | TCIA Pancreas CT | 0.00000000 |
| CAFR | 0.00003405 | ICSI MRDA Corpus | 0.000000 | OpenML | 0.000469 | TempEval-3 | 0.00007928 |
| Caltech | 0.02095834 | ICVL Hands | 0.000000 | Oxford 102 Flowers | 0.000080 | Text8 | 0.00151862 |
| CamVid | 0.00022812 | IDHP | 0.000000 | Oxford IIIT Pets | 0.000008 | The ARRAU Corpus | 0.00000959 |
| Cats and Dogs | 0.00148122 | IEMOCAP | 0.000122 | PA-100K | 0.000000 | TimeBank | 0.00053232 |
| CCGBank | 0.00002172 | IJB | 0.000530 | Par6k | 0.000000 | TIMIT | 0.00443222 |
| CelebA | 0.00244468 | ILSVRC | 0.006063 | PASCAL VOC | 0.008332 | Tox21 | 0.00029954 |
| ChaLearn | 0.00059309 | IMAGECLEF | 0.000839 | Pascal3D+ | 0.000069 | ToxCast | 0.00005048 |
| CHALL | 0.00022490 | ImageNet | 0.028748 | PATHFINDMAZES | 0.000000 | Trading Agents Competition | 0.00030921 |
| Children's Book Test | 0.00012210 | IMDb | 0.010094 | Pavia University | 0.000115 | TREC | 0.01721892 |
| CHiME | 0.00015276 | iNaturalist | 0.000089 | PCBA | 0.000113 | TreeQA | 0.00135855 |
| Chinese Poems | 0.00006816 | Indian Pines | 0.000211 | Penn Treebank | 0.009668 | TriviaQA | 0.00012031 |
| CIFAR | 0.02494334 | iPinYou | 0.000018 | PETA | 0.000678 | Tsinghua-Tencent | 0.00010600 |
| CIHP | 0.00000000 | ISBI 2012 EM Segmentation | 0.000078 | PhC-U373 | 0.000000 | Turing Test | 0.00261238 |
| Citeseer | 0.02500602 | iSEG 2017 Challenge | 0.000004 | Photo Art 50 | 0.000000 | TuSimple | 0.00002172 |
| Cityscapes | 0.00069756 | ISIC 2018 | 0.000016 | PLANNINGCOMP | 0.000000 | Twitter Dialogue | 0.00008427 |
| Click-Through Rate Prediction | 0.00097511 | ITOP | 0.000078 | PROMISE 2012 | 0.000000 | Ubuntu Dialogue | 0.00028614 |
| CliCR | 0.00000405 | IWSLT | 0.001082 | Pubmed | 0.006882 | UCF CC 50 | 0.00000809 |
| CMU-SE | 0.00001363 | JFLEG | 0.000016 | QAngaroo | 0.000016 | UCI | 0.09358595 |
| CNN / Daily Mail | 0.00039284 | JIGSAWS | 0.000302 | QASent | 0.000017 | UCI-KEEL | 0.00000809 |
| COCO | 0.00412190 | Kaggle Skin Lesion Segmentation | 0.000000 | QM9 | 0.000186 | UD | 0.00818373 |
| Cohn-Kanade | 0.00041558 | KITTI | 0.001659 | QuAC | 0.000016 | Urban100 | 0.00005708 |
| CompCars | 0.00003809 | Labeled Faces in the Wild | 0.001891 | Quasar | 0.001982 | UT Multi-view | 0.00000000 |
| COMPLEXQUESTIONS | 0.00000000 | Leeds Sports Poses | 0.000058 | Quora Question Pairs | 0.000104 | UTKFace | 0.00000000 |
| CoNLL | 0.02031269 | LexNorm | 0.000000 | R52 | 0.001746 | V-SNLI | 0.00000000 |
| CoQA | 0.00001618 | LibriSpeech | 0.000207 | R8 | 0.014510 | VggFace2 | 0.00001713 |
| Cora | 0.00826819 | LineMOD | 0.000027 | RACE | 0.019478 | Vid4 | 0.00001363 |
| CR | 0.03195242 | Loebner Prize | 0.000045 | RaFD | 0.000014 | Visual7W | 0.00020654 |
| Criteo | 0.00067999 | Long-tail emerging entities | 0.000000 | RAP | 0.002529 | VoxForge | 0.00010058 |
| CT-150 | 0.00000000 | LSUN Bedroom 256 x 256 | 0.000000 | Real-World Affective Faces | 0.000008 | WAF | 0.00068041 |
| CUB | 0.00254865 | LUNA | 0.001589 | RecipeQA | 0.000000 | WebFace | 0.00010956 |
| CUB-200-2011 | 0.00040947 | MAFA | 0.000071 | RecSys | 0.009449 | WebNLG | 0.00000809 |
| CUFS | 0.00005044 | Mandarin Chinese | 0.000233 | Reuters-21578 | 0.004667 | WebQuestions | 0.00018390 |
| CUFSF | 0.00001214 | Market 1501 | 0.000093 | Reverb | 0.000499 | Weibo NER | 0.00000000 |
| CUHK | 0.00562678 | MCTest | 0.000448 | RLCOMP | 0.000000 | WikiBio | 0.00000405 |
| DailyDialog | 0.00001618 | MediaEval | 0.000114 | Robo chat challenge | 0.000000 | WikiHop | 0.00002832 |
| DARPAGC | 0.00000000 | Medical domain | 0.003382 | Robocup | 0.004842 | Wikipedia | 0.05339900 |
| DARPARESAVE | 0.00000000 | MegaFace | 0.000164 | RotoWire | 0.000000 | WikiQA | 0.00019131 |
| DARPAUC | 0.00000000 | METR-LA | 0.000008 | RT-GENE | 0.000004 | WikiSQL | 0.00005259 |
| DBpedia | 0.00824343 | MHP | 0.000122 | RumourEval | 0.000008 | WikiText-103 | 0.00005409 |
| DCASE | 0.00033334 | Million Song Dataset | 0.001785 | RVL-CDIP | 0.000000 | WikiText-2 | 0.00018803 |
| DensePose-COCO | 0.00000809 | MIMIC-III | 0.000607 | SBD | 0.000265 | Winograd Schema Challenge | 0.00037346 |
| Dianping | 0.00014017 | Mini-ImageNet | 0.000245 | Scan2CAD | 0.000000 | Wizard-of-Oz | 0.00066401 |
| DIC HeLa | 0.00000000 | MIREX | 0.000934 | ScanNet | 0.000042 | WMT | 0.00329137 |
| DISFA | 0.00003736 | MLDoc | 0.000000 | SciTail | 0.000040 | WN18 | 0.00049114 |
| Disguised Faces in the Wild | 0.00000000 | MMI | 0.001267 | SCUT-FBP | 0.000017 | WOS | 0.00023507 |
| Douban | 0.00058997 | MNIST | 0.063154 | SearchQA | 0.000114 | WSJ | 0.00565300 |
| DRIVE | 0.04997003 | ModelNet40 | 0.000164 | Second dialogue state tracking challenge | 0.000008 | XNLI | 0.00001214 |
| DUC 2004 Task 1 | 0.00000405 | Monologue | 0.000763 | SemEval | 0.004884 | Yahoo! Answers | 0.00375964 |
| DukeMTMC-reID | 0.00002427 | MORPH | 0.002163 | SensEval | 0.000159 | YCB-Video | 0.00000000 |
| DuReader | 0.00000405 | MORPH Album2 | 0.000017 | SentEval | 0.000044 | Yelp | 0.00362348 |
| ECCV HotOrNot | 0.00000000 | MOSI | 0.000054 | Sentihood | 0.000004 | YouTube Faces | 0.00019026 |
| EMNLP 2017 | 0.00062733 | MovieLens | 0.014568 | Sequential MNIST | 0.000247 | | |
| enwiki8 | 0.00000000 | MPII | 0.000567 | ShanghaiTech | 0.000115 | | |
| NULL | NULL | MPQA | 0.002706 | ShapeNet | 0.000461 | | |

Table 7: Set of AI benchmarks and their mean intensity calculated using *AI topics*.

## Appendix E. AI Exposure Score for Studied Occupations

| ISCO 3d | Occupation | AI pct. | AI pct. (min annot.) | AI pct (max annot.) |
|---|---|---|---|---|
| 215 | Electrotechnology engineers | 1.000*** | 0.983 | 0.941 |
| 252 | Database and network professionals | 0.992*** | 1.000 | 0.874 |
| 251 | Software and applications developers and analysts | 0.983*** | 0.992 | 0.882 |
| 214 | Engineering professionals (excluding electrotechnology) | 0.975*** | 0.924 | 0.992 |
| 212 | Mathematicians, actuaries and statisticians | 0.966** | 0.824 | 0.983 |
| 351 | Information and communications technology operators | 0.958*** | 0.975 | 0.849 |
| 311 | Physical and engineering science technicians | 0.950** | 0.958 | 0.798 |
| 241 | Finance professionals | 0.941*** | 0.933 | 1.000 |
| 331 | Financial and mathematical associate professionals | 0.933*** | 0.941 | 0.958 |
| 314 | Life science technicians and related associates | 0.924** | 0.916 | 0.739 |
| 213 | Life science professionals | 0.916*** | 0.891 | 0.807 |
| 211 | Physical and earth science professionals | 0.908*** | 0.908 | 0.79 |
| 231 | University and higher education teachers | 0.899*** | 0.874 | 0.933 |
| 313 | Process control technicians | 0.891** | 0.815 | 0.672 |
| 233 | Secondary education teachers | 0.882*** | 0.866 | 0.916 |
| 216 | Architects, planners, surveyors and designers | 0.874*** | 0.798 | 0.866 |
| 613 | Mixed crop and animal producers | 0.866 | 0.899 | 0.403 |
| 431 | Numerical clerks | 0.857*** | 0.807 | 0.756 |
| 352 | Telecommunications and broadcasting technicians | 0.849* | 0.882 | 0.538 |
| 413 | Keyboard operators | 0.840 | 0.950 | 0.420 |
| 242 | Administration professionals | 0.832*** | 0.832 | 0.891 |
| 334 | Administrative and specialised secretaries | 0.824*** | 0.748 | 0.824 |
| 243 | Sales, marketing and public relations professionals | 0.815** | 0.731 | 0.899 |
| 264 | Authors, journalists and linguists | 0.807** | 0.849 | 0.689 |
| 411 | General office clerks | 0.798** | 0.857 | 0.664 |
| 133 | Information and communications technology services | 0.790 | 0.513 | 0.924 |
| 122 | Sales, marketing and development managers | 0.782 | 0.529 | 0.966 |
| 412 | Secretaries (general) | 0.773*** | 0.756 | 0.714 |
| 232 | Vocational education teachers | 0.765** | 0.672 | 0.815 |
| 121 | Business services and administration managers | 0.756 | 0.496 | 0.95 |
| 732 | Printing trades workers | 0.748 ** | 0.773 | 0.597 |
| 132 | Manufacturing, mining, construction, and distribution managers | 0.739 | 0.412 | 0.975 |
| 261 | Legal professionals | 0.731*** | 0.840 | 0.782 |
| 831 | Locomotive engine drivers and related workers | 0.723 | 0.966 | 0.252 |
| 221 | Medical doctors | 0.714*** | 0.723 | 0.723 |
| 234 | Primary school and early childhood teachers | 0.706*** | 0.697 | 0.731 |
| 441 | Other clerical support workers | 0.697*** | 0.664 | 0.655 |
| 742 | Electronics and telecommunications installers ... | 0.689* | 0.782 | 0.471 |
| 621 | Forestry and related workers | 0.681** | 0.647 | 0.496 |
| 321 | Medical and pharmaceutical technicians | 0.672** | 0.739 | 0.529 |
| 262 | Librarians, archivists and curators | 0.664** | 0.79 | 0.639 |
| 315 | Ship and aircraft controllers and technicians | 0.655*** | 0.706 | 0.588 |
| 333 | Business services agents | 0.647** | 0.63 | 0.765 |

| ISCO 3d | Occupation | AI pct. | AI pct. (min annot.) | AI pct (max annot.) |
|---|---|---|---|---|
| 432 | Material-recording and transport clerks | 0.639* | 0.345 | 0.681 |
| 263 | Social and religious professionals | 0.63*** | 0.714 | 0.697 |
| 812 | Metal processing and finishing plant operators | 0.622 | 0.765 | 0.118 |
| 134 | Professional services managers | 0.613 | 0.328 | 0.908 |
| 343 | Artistic, cultural and culinary associate professionals | 0.605*** | 0.613 | 0.504 |
| 235 | Other teaching professionals | 0.597** | 0.555 | 0.706 |
| 332 | Sales and purchasing agents and brokers | 0.588* | 0.538 | 0.84 |
| 741 | Electrical equipment installers and repairers | 0.58** | 0.655 | 0.429 |
| 752 | Wood treaters, cabinet-makers and related trades | 0.571 | 0.639 | 0.059 |
| 111 | Legislators and senior officials | 0.563* | 0.454 | 0.773 |
| 723 | Machinery mechanics and repairers | 0.555** | 0.622 | 0.395 |
| 112 | Managing directors and chief executives | 0.546 | 0.429 | 0.832 |
| 817 | Wood processing and papermaking plant operators | 0.538 | 0.681 | 0.092 |
| 341 | Legal, social and religious associate professionals | 0.529*** | 0.58 | 0.613 |
| 813 | Chemical and photographic products plant and m... | 0.521 | 0.605 | 0.227 |
| 335 | Regulatory government associate professionals | 0.513*** | 0.571 | 0.563 |
| 722 | Blacksmiths, toolmakers and related trades workers | 0.504* | 0.689 | 0.336 |
| 721 | Sheet and structural metal workers, moulders a... | 0.496 | 0.588 | 0.126 |
| 325 | Other health associate professionals | 0.487*** | 0.521 | 0.487 |
| 611 | Market gardeners and crop growers | 0.479** | 0.395 | 0.311 |
| 612 | Animal producers | 0.471* | 0.546 | 0.185 |
| 322 | Nursing and midwifery associate professionals | 0.462* | 0.269 | 0.555 |
| 226 | Other health professionals | 0.454*** | 0.462 | 0.571 |
| 816 | Food and related products machine operators | 0.445* | 0.336 | 0.193 |
| 225 | Veterinarians | 0.437** | 0.361 | 0.605 |
| 265 | Creative and performing artists | 0.429*** | 0.403 | 0.521 |
| 818 | Other stationary plant and machine operators | 0.42* | 0.378 | 0.16 |
| 422 | Client information workers | 0.412** | 0.563 | 0.58 |
| 143 | Other services managers | 0.403 | 0.235 | 0.748 |
| 712 | Building finishers and related trades workers | 0.395*** | 0.387 | 0.361 |
| 421 | Tellers, money collectors and related clerks | 0.387* | 0.445 | 0.647 |
| 511 | Travel attendants, conductors and guides | 0.378*** | 0.487 | 0.445 |
| 142 | Retail and wholesale trade managers | 0.370 | 0.193 | 0.857 |
| 516 | Other personal services workers | 0.361 | 0.244 | 0.63 |
| 753 | Garment and related trades workers | 0.353* | 0.437 | 0.151 |
| 224 | Paramedical practitioners | 0.345*** | 0.42 | 0.294 |
| 312 | Mining, manufacturing and construction supervisors | 0.336* | 0.202 | 0.546 |
| 815 | Textile, fur and leather products machine operators | 0.328 | 0.597 | 0.025 |
| 811 | Mining and mineral processing plant operators | 0.319** | 0.286 | 0.437 |
| 731 | Handicraft workers | 0.311*** | 0.37 | 0.345 |
| 324 | Veterinary technicians and assistants | 0.303* | 0.471 | 0.143 |
| 541 | Protective services workers | 0.294*** | 0.353 | 0.412 |
| 531 | Child care workers and teachers' aides | 0.286** | 0.218 | 0.378 |
| 754 | Other craft and related workers | 0.277*** | 0.277 | 0.387 |
| 821 | Assemblers | 0.269 | 0.479 | 0.042 |
| 832 | Car, van and motorcycle drivers | 0.261*** | 0.252 | 0.353 |

| ISCO 3d | Occupation | AI pct. | AI pct. (min annot.) | AI pct (max annot.) |
|---|---|---|---|---|
| 342 | Sports and fitness workers | 0.252** | 0.294 | 0.479 |
| 222 | Nursing and midwifery professionals | 0.244* | 0.176 | 0.454 |
| 814 | Rubber, plastic and paper products machine operators | 0.235 | 0.504 | 0.109 |
| 833 | Heavy truck and bus drivers | 0.227*** | 0.303 | 0.303 |
| 931 | Mining and construction labourers | 0.218*** | 0.261 | 0.134 |
| 713 | Painters, building structure cleaners and related trades | 0.21** | 0.319 | 0.101 |
| 532 | Personal care workers in health services | 0.202*** | 0.151 | 0.235 |
| 932 | Manufacturing labourers | 0.193** | 0.101 | 0.277 |
| 835 | Ships' deck crews and related workers | 0.185** | 0.118 | 0.319 |
| 711 | Building and related trades in construction | 0.176*** | 0.168 | 0.218 |
| 512 | Cooks | 0.168*** | 0.210 | 0.210 |
| 834 | Mobile plant operators | 0.16** | 0.185 | 0.05 |
| 524 | Other sales workers | 0.151** | 0.227 | 0.370 |
| 515 | Building and housekeeping supervisors | 0.143*** | 0.143 | 0.176 |
| 751 | Food processing and related trades workers | 0.134*** | 0.160 | 0.261 |
| 962 | Other elementary workers | 0.126* | 0.109 | 0.269 |
| 141 | Hotel and restaurant managers | 0.118 | 0.084 | 0.622 |
| 521 | Street and market salespersons | 0.109 | 0.076 | 0.513 |
| 622 | Fishery workers, hunters and trappers | 0.101*** | 0.134 | 0.076 |
| 522 | Shop salespersons | 0.092 | 0.126 | 0.462 |
| 514 | Hairdressers, beauticians and related workers | 0.084** | 0.311 | 0.286 |
| 933 | Transport and storage labourers | 0.076*** | 0.067 | 0.168 |
| 523 | Cashiers and ticket clerks | 0.067** | 0.092 | 0.202 |
| 921 | Agricultural, forestry and fishery labourers | 0.059*** | 0.042 | 0.008 |
| 961 | Refuse workers | 0.050*** | 0.034 | 0.017 |
| 513 | Waiters and bartenders | 0.042** | 0.059 | 0.244 |
| 911 | Domestic, hotel and office cleaners and helpers | 0.034*** | 0.025 | 0.084 |
| 941 | Food preparation assistants | 0.025*** | 0.050 | 0.034 |
| 912 | Vehicle, window, laundry and other hand cleani... | 0.017*** | 0.017 | 0.067 |
| 952 | Street vendors (excluding food) | 0.008 * | 0.008 | 0.328 |

Table 8: AI exposure score (percentiles) by occupation

Stars denote level of certainty: $***$ interpercentile range <= 0.129, $**$ interpercentile range <= 0.249, and $*$ interpercentile range <= 0.369. Levels of certainty were picked based on an elbow method. Min annotations denote ability-task annotations reduced by one standard-deviation. Max annotations denote ability-task annotations increased by one standard deviation

## Appendix F. Further Results

Figure 9 shows scores of matrix $\mathbf{W}(119 \times 14)$ for nine selected ISCO-3 occupations: general office clerks, shop salespersons, cleaners and helpers, medical doctors, personal care workers in health services, primary school and early childhood teachers, heavy truck and bus drivers, waiters and bartenders, building and related trades in construction. That is, the figure shows for each of the nine selected occupations the relevance of each cognitive ability relative to the other cognitive abilities. In each subfigure we highlight one of the nine selected occupations. As above, each subfigure is divided into the categories according to the objects they deal with: *people, ideas,*

and *things*. As was expected from the narrow distributions shown in Figure 5, the selected occupations show similar relevance profiles.

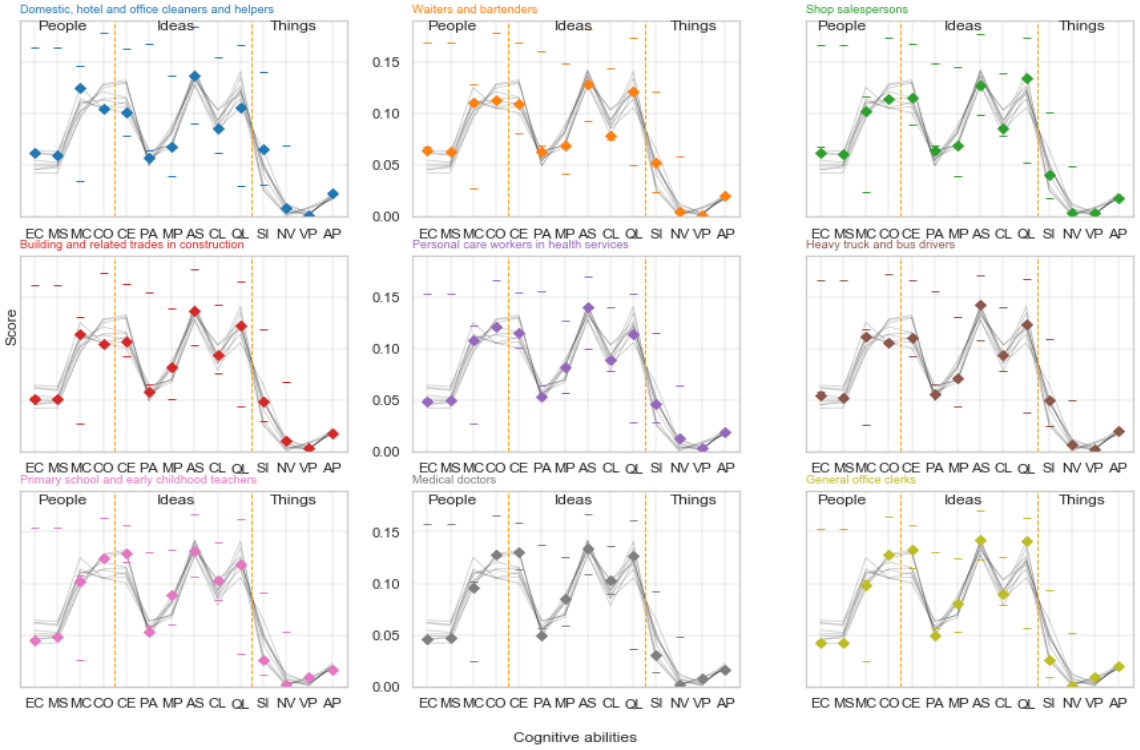



Figure 9: Ability-specific scores of cognitive abilities for selected occupations. Coloured markers represent the highlighted occupation. Coloured dashes represent intervals due to perturbation of ability-task assignments. Grey lines visualise the relevance scores for the other eight occupations.

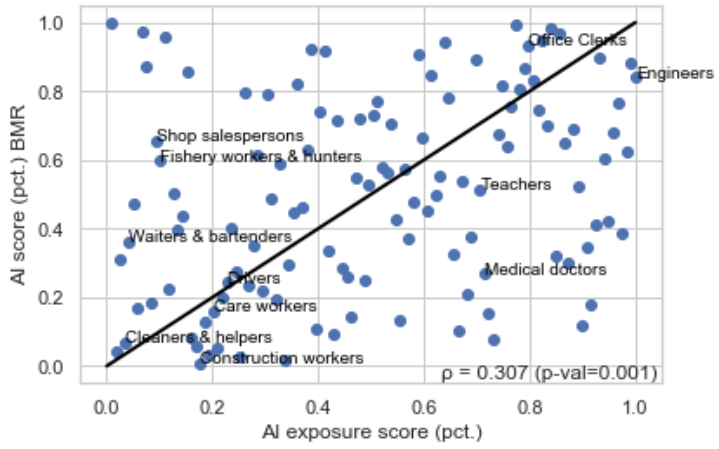
Furthermore, the figure shows that medical doctors, teachers and office clerks have high intensity scores for most abilities in the category *ideas*. These occupations also exhibit less pronounced scores for *sensorimotor interaction* (SI). In contrast, heavy truck and bus drivers, and workers in building and related trades in construction have lower intensity levels for abilities dealing with *people* and *ideas* (except for *attention and search*(AS), which is high for every occupation) but higher intensity levels for *sensorimotor interaction* (SI). Furthermore, domestic, hotel and office cleaners have the highest relevance score for *sensorimotor interaction* (SI) while also requiring high levels of abilities that deal with *people* (except for *communication*(CO). Finally, shop salespersons and waiters and bartenders have relatively high levels for the *people* cognitive abilities, while the levels for the people-related abilities *emotional control* (EC) and *social interaction* (MS) are relatively low for general office clerks, medical doctors and teachers (although these occupations require high levels of *communication* (CO)). Note that this does not mean that shop salespersons require higher *emotional control* and *social interaction* than teachers. Instead, it means that social abilities for teachers are on average a less relevant part of their occupation in relation to the relevance of other abilities (such as *comprehension and expression* (CE) and *com-*

*munication*(CO)), than for shop salespersons.

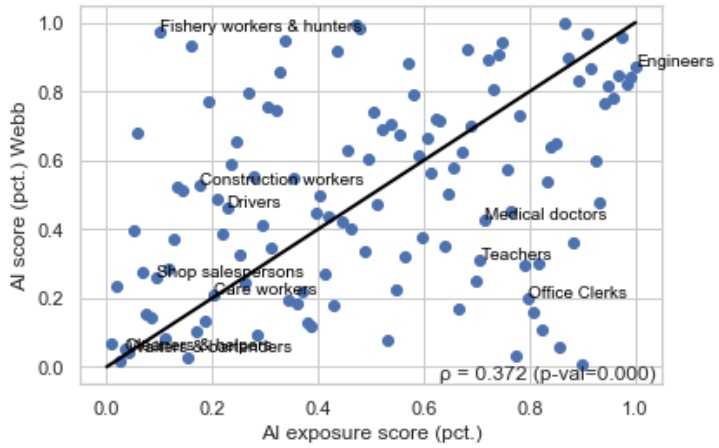
This figure also shows the uncertainty that derives from the manual annotation of cognitive abilities to tasks (see section 4.1) with coloured dashes.[35] Indeed, the added uncertainty yields a large ranges that overlap. Nevertheless, these intervals are of equal size across occupations, and the relations between the maximum potential score and minimum potential score remain the same across abilities and occupations. Therefore, this should have a limited effect on the ranking of occupations by potential AI exposure.

For its part, Figure 10 shows the scatterplots between the percentiles of the AI exposure score in this study and the percentiles of the other AI exposure scores as well as the corresponding Spearman rank correlations (ρ). Each of the score coming from the three studies is significantly correlated (with p-values ≤ 0.001) with the AI exposure score in our study, although at relatively low levels: $\rho = 0.307$ for the correlation with BMR, $\rho = 0.372$ for the correlation with Webb and a higher $\rho = 0.455$ with the FRS, which is based on an approach closest to the one taken in this paper. The reason for these differences may be rooted in the different sources for the measures of AI capabilities and consequently the different focuses set on the measurement. For instance, the much higher FRS score for drivers may be due to the fact that Felten et al. (2018) rely on AI benchmarks on the EFF platform which has a strong focus on perception benchmarks or the difference may occur because the task framework used in this paper does not explicitly cover navigation tasks. Nevertheless, from the highlighted occupations we can see that the scores of this study are always very similar to the other scores for cleaners & helpers, waiters & bartenders, care workers and electrotechnology engineers. Thus, despite the different sources and methodology of constructing AI exposure scores we find significant correlations.

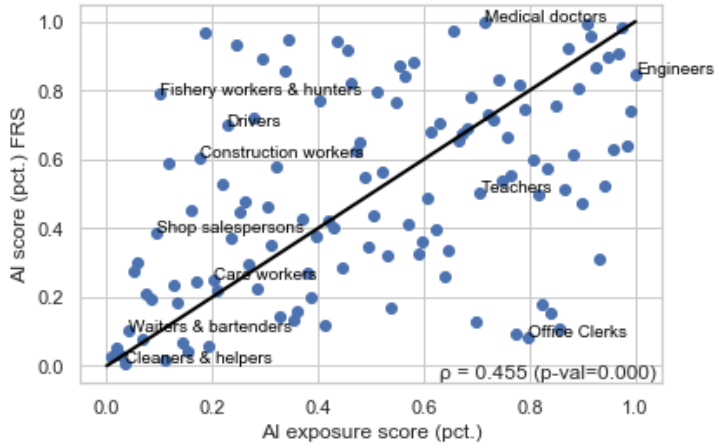
35. More precisely, we allow for uncertainty by perturbing the values of the task-ability correspondence annotation matrix (Ω) by one standard deviation upwards and downwards.




(a) (Brynjolfsson et al., 2018)




(b) (Webb, 2020)




(c) (Felten et al., 2018)

Figure 10: Correlation plots with other AI exposure scores. Black line represents the 45 degree line (ρ = 1).

## References

Acemoglu, D. (2002). Technical change, inequality, and the labor market. *Journal of Economic Literature 40*(1), 7–72.

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

Acemoglu, D. and P. Restrepo (2018). The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. *American Economic Review 108*(6), 1488–1542.

Adams, S. S., I. Arel, J. Bach, R. Coop, R. Furlan, B. Goertzel, J. S. Hall, A. Samsonovich, M. Scheutz, M. Schlesinger, S. C. Shapiro, and J. Sowa (2012). Mapping the landscape of human-level artificial general intelligence. *AI Magazine 33*(1), 25–42.

Adams, S. S., G. Banavar, and M. Campbell (2016). I-athlon: Towards a multi-dimensional Turing test. *AI Magazine 37*(1), 78–84.

Agrawal, A., J. Horton, N. Lacetera, and E. Lyons (2015). Digitization and the contract labor market: A research agenda. In *Economic Analysis of the Digital Economy*, pp. 219–250. University of Chicago Press.

Archibugi, D. (1992). Patenting as an Indicator of Technological Innovation: A Review. *Science and Public Policy 19*(6), 357–368.

Arntz, M., T. Gregory, and U. Zierahn (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. (2014). Polanyi's Paradox and the Shape of Employment Growth. Working Paper 20485, National Bureau of Economic Research.

Autor, D. (2015). Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives 29*(3), 3–30.

Autor, D. and D. Dorn (2013). The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market. *American Economic Review 103*(5), 1553–97.

Autor, D. H. (2013). The "Task Approach" to Labor Markets: An Overview. Working Paper 18711, National Bureau of Economic Research.

Autor, D. H. and M. J. Handel (2013). Putting Tasks to the Test: Human Capital, Job Tasks, and Wages. *Journal of Labor Economics 31*(S1), S59–S96.

Autor, D. H., F. Levy, and R. J. Murnane (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics 118*(4), 1279–1333.

Bessen, J. (2019). Automation and Jobs: When Technology Boosts Employment. *Economic Policy 34*(100), 589–626.

Bogliacino, F., M. Piva, and M. Vivarelli (2012). R&D and Employment: An Application of the LSDVC Estimator Using European Microdata. *Economics Letters 116*(1), 56–59.

Bottou, L., C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. D. Jackel, Y. LeCun, U. A. Müller, E. Säckinger, P. Y. Simard, et al. (1994). Comparison of Classifier Methods: A Case Study in Handwritten Digit Recognition. In *International Conference on Pattern Recognition*, pp. 77–77. IEEE Computer Society Press.

Brown, N. and T. Sandholm (2019). Superhuman AI for multiplayer poker. *Science 365*(6456), 885–890.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Brundage, M. (2016). Modeling Progress in AI. In *AI, Ethics, and Society, Papers from the 2016 AAAI Workshop, Phoenix, Arizona, USA, February 13, 2016*.

Brynjolfsson, E. and T. Mitchell (2017). What can machine learning do? Workforce implications. *Science 358*(6370), 1530–1534.

Brynjolfsson, E., T. Mitchell, and D. Rock (2018). What Can Machines Learn, and What Does It Mean for Occupations and the Economy? In *AEA Papers and Proceedings*, Volume 108, pp. 43–47.

Brynjolfsson, E., D. Rock, and C. Syverson (2018). Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Brynjolfsson, E., D. Rock, and C. Syverson (2021). The Productivity J-Curve: How Intangibles Complement General Purpose Technologies. *American Economic Journal: Macroeconomics 13*(1), 333–72.

Buchanan, B. G., J. Eckroth, and R. Smith (2013). A Virtual Archive for the History of AI. *AI Magazine 34*(2), 86–86.

Calvino, F. and M. E. Virgillito (2018). The Innovation-Employment Nexus: A Critical Survey of Theory and Empirics. *Journal of Economic Surveys 32*(1), 83–117.

Campbell, M., A. J. Hoane, and F. Hsu (2002). Deep Blue. *Artificial Intelligence 134*(1-2), 57 – 83.

Carroll, J. B. et al. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press.

Ciarli, T., A. Marzucchi, E. Salgado, and M. Savona (2018). The Effect of R&D Growth on Employment and Self-Employment in Local Labour Markets. SPRU Working Paper Series 2018-08, SPRU - Science Policy Research Unit, University of Sussex Business School.

Claveria, O., E. Monte, and S. Torra (2019). Economic Uncertainty: A Geometric Indicator of Discrepancy Among Experts' Expectations. *Social Indicators Research*, 1–20.

Coad, A. and R. Rao (2011). The Firm-Level Employment Effects of Innovations in High-Tech US Manufacturing Industries. *Journal of Evolutionary Economics 21*(2), 255–283.

Cockburn, I. M., R. Henderson, and S. Stern (2018). The Impact of Artificial Intelligence on Innovation. Working Paper 24449, National Bureau of Economic Research.

Coxeter, H. S. M. (1961). *Introduction to Geometry*. John Wiley & Sons, New York, London.

Dalkey, N. and O. Helmer (1963). An Experimental Application of the Delphi Method to the Use of Experts. *Management Science 9*(3), 458–467.

Davis, E. and G. Marcus (2015). Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence. *Communications of the ACM 58*(9), 92–103.

Deming, D. J. (2017). The Growing Importance of Social Skills in the Labor Market. *The Quarterly Journal of Economics 132*(4), 1593–1640.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). Imagenet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee.

Dosi, G. (1988). Sources, Procedures, and Microeconomic Effects of Innovation. *Journal of Economic Literature*, 1120–1171.

Felten, E. W., M. Raj, and R. Seamans (2018). A Method to Link Advances in Artificial Intelligence to Occupational Abilities. In *AEA Papers and Proceedings*, Volume 108, pp. 54–57.

Fernandez-Macias, E. and M. Bisello (2020). A Taxonomy of Tasks for Assessing the Impact of New Technologies on Work. Technical report, Joint Research Centre (Seville site).

Fernández-Macías, E., E. Gómez, J. Hernández-Orallo, B. S. Loe, B. Martens, F. Martínez-Plumed, and S. Tolan (2018). A Multidisciplinary Task-Based Perspective for Evaluating the Impact of AI Autonomy and Generality on the Future of Work. *arXiv preprint arXiv:1807.02416*.

Fernández-Macías, E., J. Hurley, and M. Bisello (2016). *What Do Europeans Do at Work?: A Task-based Analysis*. Publication Office of the European Union.

Ferrucci, D. A. (2012). Introduction to "This is Watson". *IBM Journal of Research and Development 56*(3.4), 1–1.

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 M. A. Osborne (2017). The Future of Employment: How Susceptible are Jobs to Computerisation? *Technological Forecasting and Social Change 114*, 254–280.

Goos, M., A. Manning, and A. Salomons (2009). Job Polarization in Europe. *American Economic Review 99*(2), 58–63.

Goos, M., A. Manning, and A. Salomons (2014). Explaining Job Polarization: Routine-Biased Technological Change and Offshoring. *American Economic Review 104*(8), 2509–26.

Grace, K., J. Salvatier, A. Dafoe, B. Zhang, and O. Evans (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research 62*, 729–754.

Gries, T. and W. Naudé (2020). Artificial Intelligence, Income Distribution and Economic Growth. IZA Discussion Papers 13606, Institute of Labor Economics (IZA).

Griliches, Z. (1998). Patent Statistics as Economic Indicators: A Survey. In *R&D and Productivity: The Econometric Evidence*, pp. 287–343. University of Chicago Press.

Hendrycks, D., C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt (2020). Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*.

Henighan, T., J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, et al. (2020). Scaling Laws for Autoregressive Generative Modeling. *arXiv preprint arXiv:2010.14701*.

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.

Hernández-Orallo, J., M. Baroni, J. Bieger, N. Chmait, D. L. Dowe, K. Hofmann, F. Martínez-Plumed, C. Strannegård, and K. R. Thórisson (2017). A New AI Evaluation Cosmos: Ready to Play the Game? *AI Magazine 38*(3).

Hernández-Orallo, J. and K. Vold (2019). AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 507–513.

Jaffe, A. and G. de Rassenfosse (2017). Patent Citation Data in Social Science Research: Overview and Best Practices. *Journal of the Association for Information Science and Technology 68*(6), 1360–1374.

Keith, T. Z. and M. R. Reynolds (2010). Cattell–Horn–Carroll Abilities and Cognitive Tests: What We've Learned from 20 Years of Research. *Psychology in the Schools 47*(7), 635–650.

Krizhevsky, A., G. Hinton, et al. (2009). Learning Multiple Layers of Features from Tiny Images. Technical report, Citeseer.

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick (2014). Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, pp. 740–755. Springer.

Machado, M. C., M. G. Bellemare, E. Talvitie, J. Veness, M. Hausknecht, and M. Bowling (2018). Revisiting the Arcade Learning Environment: Evaluation Protocols and Open Problems for General Agents. *Journal of Artificial Intelligence Research 61*, 523–562.

Manning, C. D., C. D. Manning, and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. MIT press.

Martínez-Plumed, F., S. Avin, M. Brundage, A. Dafoe, S. Ó. hÉigeartaigh, and J. Hernández-Orallo (2018). Accounting for the Neglected Dimensions of AI Progress. *arXiv preprint arXiv:1806.00610*.

Martínez-Plumed, F., E. Gómez, and J. Hernández-Orallo (2020a). Tracking AI: The Capability is (Not) Near. In *Proceedings of the Twenty-six European Conference on Artificial Intelligence*. IOS Press.

Martínez-Plumed, F., E. Gómez, and J. Hernández-Orallo (2020b). Tracking the Evolution of AI: The AIcollaboratory. In *Proceedings of the 1st International Workshop: Evaluating Progress in Artificial Intelligence (EPAI 2020)*.

Martinez-Plumed, F. and J. Hernandez-Orallo (2018). Dual Indicators to Analyse AI Benchmarks: Difficulty, Discrimination, Ability and Generality. *IEEE Transactions on Games*, 1–1.

Martínez-Plumed, F., S. Tolan, A. Pesole, J. Hernández-Orallo, E. Fernández-Macías, and E. Gómez (2020). Does AI Qualify for the Job? A Bidirectional Model Mapping Labour and AI Intensities. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 94–100.

Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis (2015). Human-Level Control Through Deep Reinforcement Learning. *Nature 518*, 529–533.

Mokyr, J., C. Vickers, and N. L. Ziebarth (2015). The History of Technological Anxiety and the Future of Economic Growth: Is This Time Different? *Journal of Economic Perspectives 29*(3), 31–50.

Müller, V. C. and N. Bostrom (2014). Future Progress in Artificial Intelligence: A Poll Among Experts. *AI Matters 1*(1), 9–11.

Müller, V. C. and N. Bostrom (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In *Fundamental Issues of Artificial Intelligence*, pp. 555–572. Springer.

Nedelkoska, L. and G. Quintini (2018). Automation, Skills Use and Training. *OECD Social, Employment and Migration Working Papers* (202).

Nuti, S. V., B. Wayda, I. Ranasinghe, S. Wang, R. P. Dreyer, S. I. Chen, and K. Murugiah (2014). The Use of Google Trends in Health Care Research: A Systematic Review. *PloS one 9*(10), e109583.

Polanyi, M. (1966). The Logic of Tacit Inference. *Philosophy 41*(155), 1–18.

Preis, T., H. S. Moat, and H. E. Stanley (2013). Quantifying Trading Behavior in Financial Markets Using Google Trends. *Scientific Reports 3*, 1684.

Purves, C., C. Cangea, and P. Veličković (2019). The PlayStation Reinforcement Learning Environment (PSXLE). *arXiv preprint arXiv:1912.06101*.

Rajpurkar, P., R. Jia, P. Liang, and . (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. *arXiv preprint arXiv:1806.03822*.

Saari, D. G. (2008). Complexity and the Geometry of Voting. *Mathematical and Computer Modelling 48*(9-10), 1335–1356.

Schaie, K. W. (2010). Primary Mental Abilities. In I. B. Weiner and W. E. Craighead (Eds.), *Corsini Encyclopedia of Psychology*, pp. 1286–1288. Wiley.

Shoham, Y., R. Perrault, E. Brynjolfsson, J. Clark, J. Manyika, J. C. Niebles, T. Lyons, J. Etchemendy, and Z. Bauer (2018). The AI Index 2018 Annual Report. *AI Index Steering Committee, Human-Centered AI Initiative, Stanford University. 202018*.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature 529*(7587), 484.

Silver, D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. (2017). Mastering the game of Go without human knowledge. *Nature 550*(7676), 354–359.

Vinyals, O. et al. (2017). Starcraft II: A New Challenge for Reinforcement Learning. *arXiv preprint arXiv:1708.04782*.

Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman (2019). Superglue: A Stickier Benchmark for General-Purpose Language Understanding Systems. *arXiv preprint arXiv:1905.00537*.

Wasserman, E. A. and T. R. Zentall (2006). *Comparative Cognition: Experimental Explorations of Animal Intelligence*. Oxford University Press.

Webb, M. (2020). The Impact of Artificial Intelligence on the Labor Market. *Department of Economics, Stanford University, Stanford, California*.

Whiteson, S., B. Tanner, M. E. Taylor, and P. Stone (2011). Protecting Against Evaluation Overfitting in Empirical Reinforcement Learning. In *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pp. 120–127. IEEE.