1. New approaches to understanding the impact of computers on work and education

Nóra Révai, OECD

Mila Staneva, OECD

Abel Baret, OECD

This chapter describes the background and purpose of the OECD's *Artificial Intelligence and the Future of Skills* project, which is developing an approach to assessing the capabilities of artificial intelligence (AI) and robotics and their impact on education and work. This report represents the project's first step to identify the capabilities to assess and the tests to use for the assessment. The chapter provides an overview of the approaches applied to date to predict the impact of technology on the future of work. It sets out a new approach and presents the project's stages of developing a sound methodology for a systematic assessment of AI capabilities in the future. The chapter ends by presenting the structure of this report.

Introduction

Policy interest in the impact of artificial intelligence (AI) has sprung up in the past few decades as AI technologies are developing and being integrated into more and more aspects of life. A deeper and more precise understanding of this impact for the economy and society is fundamental for strategic planning in various policy areas. With regard to employment and education, this understanding can provide the basis for realistic scenarios about how jobs and skill demand will be redefined in the next decades. It can also demonstrate how the education system needs to be reshaped to prepare today's students for these possible futures.

However, this understanding of the impacts of AI and robotics fundamentally rests on an understanding of the technology's capabilities. What can AI and robotics do and what can they not do? How do the capabilities of AI and robotics compare to those of humans?

The OECD's Centre for Educational Research and Innovation (CERI) launched the *Artificial Intelligence and the Future of Skills* (AIFS) project in 2019 to address the questions above. The project builds on pilot work carried out in 2016 that explores AI capabilities with respect to literacy, numeracy and problem-solving skills using the OECD's Survey of Adult Skills. The project aims to develop a new set of measures to serve as a foundation for research and policy on how AI and robotics will transform skill demand and educational requirements in the decades ahead. It addresses the following concrete questions:

- What human capabilities will be too difficult for AI and robotics to reproduce over the next few decades?
- What education and training will be needed to allow most people to develop some work-related capabilities that are beyond the capabilities of AI and robotics?

Studies that have attempted to gauge the impact of computer capabilities on employment, skill demand and education demonstrate that predictions on work and society are by no means straightforward. Technological development affects the labour market in diverse ways that are sometimes hard to predict. It usually involves the transformation of jobs and tasks rather than their full replacement by machines.

Despite these complexities, an understanding of that transformation must begin with an understanding of the computer capabilities themselves. This can provide a basis for reflecting on potential transformations. This report does *not* discuss the implications of technological change, or even reflect on how to identify those implications. Rather, it aims to build a methodology that can provide valuable and robust data for policy makers and researchers who want to consider those transformations from a solid understanding of the technology itself.

Accordingly, the first stage of the project focuses on constructing *valid*, *reliable* and *meaningful* measures of AI capabilities. First, to be valid, measures must not mislead and indeed assess capabilities of AI. Second, to be reliable, measures must ensure consistency over time. To that end, they must rely on recognised experts, and a transparent and robust process that is reproducible. Reliability also involves addressing convergences and divergences of experts' judgements transparently and appropriately (inter-rater reliability), and using consistent items for measurement. Validity and reliability make the measures credible, ensuring that the measurement avoids basic methodological pitfalls. Third, for measures to be meaningful, particularly to the policy community, the constructs and comparisons should enable decision makers without AI expertise to understand likely implications. For this to happen, the set of measures should be comprehensive, covering the full spectrum of relevant capabilities. A straightforward comparison with human capabilities would help interpret constructs and help produce meaningful measures.

Such a set of measures requires scientific thoroughness and broad partnerships to be effective and comprehensive. Consequently, the project dedicates substantial effort to build a robust methodology and involve a wide range of experts from around the world. Developing the AIFS approach will take place over six years, which began with a planning process in 2019.

This first volume explores the methodology for the project. It is a technical report, which provides the outcomes of an expert workshop held in October 2020 on "Skills and Tests". This initial workshop provided the direction for pilot work in 2021 on different types of assessment tasks. The pilot will be followed by an initial systematic assessment in 2022 and 2023 across the full range of capabilities. An analysis of the potential implications for work and education will be produced in 2024. The project will conclude with a proposed approach for a regular programme to update the assessments.

This first chapter of the technical report situates the project in the broader literature on evaluating the progress of computer technology and its impact on the world of work. It starts with a snapshot of the broader effort of the OECD to gauge the impact of AI. Next, it discusses the various methodologies adopted to date and their challenges. It then presents the pilot work underpinning the AIFS project, showing how a new approach can address some of the methodological caveats and gaps. The report structure is presented at the end of this chapter.

OECD's work on the impact of artificial intelligence

Over the past decade, advances in big data, computational power, storage capacity and algorithmic techniques have dramatically accelerated the development and deployment of AI systems. The OECD has been increasingly engaged in supporting countries to understand this technological development. A major achievement was the adoption of the OECD Principles on Artificial Intelligence in May 2019, which sets international standards for the responsible stewardship of trustworthy AI (OECD, 2019[1]). In 2020, the OECD launched the AI Policy Observatory, which brings together information, analysis and supports dialogue to shape and share AI policies (OECD, 2020[2]).

A recent cross-directorate effort of the OECD is the AI programme on Work, Innovation, Productivity and Skills (AI-WIPS) supported by the German Federal Ministry of Labour and Social Affairs. AI-WIPS incorporates several streams of work to provide a comprehensive analysis of the different aspects of AI and their implications for society. The AIFS project represents the "Assessing AI and robotics capabilities" work stream. In parallel, work has started on building a framework for classifying AI systems and mapping their development in various fields (Baruffaldi et al., 2020[3]; OECD, 2019[4]). The OECD has been contributing to analysing the impact of AI on workforce skills in the past few years. As part of this work, analyses have been conducted on the extent to which machines can automate jobs and substitute for workers (Arntz, Gregory and Zierahn, 2016[5]; Nedelkoska and Quintini, 2018[6]). A more recent publication reviews literature on the impact of AI on employment, wages, the work environment and the ways in which AI transforms jobs and skill needs (Lane and Saint-Martin, 2021[7]). As the impact of technology on work and society also depends on the speed of its development and diffusion, the OECD is also working on assessing the speed of AI diffusion (Nakazato and Squicciarini, 2021[8]). Finally, the organisation has been facilitating policy and societal dialogue that brings together experts, researchers, policy makers, social partners and civil society to discuss and contribute to AI-related topics.

In the domain of education, CERI engages in several aspects of understanding the impact of technology on education (Vincent-Lancrin and van der Vlies, 2020[9]; van der Vlies, 2020[10]). Work on educational innovation explores the uses of digital devices and software for enhancing learning inside and outside the classroom [see also (Verhagen, forthcoming[11])]. This work extends to understanding how data – whether collected in formal educational settings or through other means – can be used to personalise learning, improve people's educational experience, and inform decision making and policies in education.

Box 1.1. Artificial intelligence: Definition, use cases, scope

There is no commonly agreed definition of AI systems (OECD, 2019[4]). While machine learning has become popular (see Chapter 12 for more information), computer scientists stress that the understanding of AI should extend beyond this technique. The OECD's AI Experts Group (AIGO) defines an AI system as:

> a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. It uses machine and/or human-based inputs to perceive real and/or virtual environments; abstract such perceptions into models (in an automated manner e.g. with machine learning (ML) or manually); and use model inference to formulate options for information or action. AI systems are designed to operate with varying levels of autonomy *(OECD, 2019, p. 15[4])*.

The ability to "make predictions, recommendations or decisions" makes AI applicable in various tasks and domains. A recent OECD report presents a number of use cases across a range of areas (OECD, 2019[4]). In health care, for example, AI is used to make more accurate and faster diagnoses of diseases than humans. In e-commerce, AI is applied to recommend products that better fit the needs of potential buyers. In finance, AI is used to predict the credibility of loan applicants more accurately. AI can also automate numerous other tasks where the role of prediction is less obvious. One example is autonomous driving, where cars rely on AI to anticipate the right trajectories and manoeuvres. Each of these areas has important limits with respect to AI's capabilities, but the wide applicability in different economic sectors makes AI a "general-purpose technology" (OECD, 2019[4]). Like the steam engine and electricity, AI has the potential to raise productivity across vast parts of the economy (Bresnahan and Trajtenberg, 1992[12]).

Although AI is being applied to increasingly more areas, it still cannot perform the full range of tasks of humans and lacks some basic human skills. Therefore, one way to describe the types of AI that currently exist is that they represent artificial narrow intelligence, meaning that current AI systems are designed to perform specific, narrowly defined tasks (OECD, 2017[13]). This state-of-the-art is contrasted to a hypothetical artificial general intelligence (AGI) (OECD, 2017[13]). In AGI, machines are human-like in how they can abstract, generalise, perceive, judge, create and make decisions. Such skills are still out of reach for AI systems (OECD, 2019[4]).

Parallel to progress in AI, robot technology has continued to advance (OECD, 2021[14]). Indeed, both these technological developments are intertwined as many new AI applications involve the sensory and motor control capabilities that are fundamental to robotics. In addition, both AI and robotic technologies enable the automation of tasks typically executed by humans. As a result, they have a similar impact on the world of work. For all these reasons, advancements in AI and robotics and their implications for the future of work are commonly studied together in the literature. The current report is centred on AI but encompasses robotic technologies as they are also likely to affect work and the future demand of skills.

Attempts to measure artificial intelligence capabilities and impact

There is general agreement that AI is a major breakthrough technology that will transform the economy and society (see Box 1.1 for definitions, use cases and the scope of AI). However, to unravel this impact, studies must first understand what computers can and cannot do. Most of the work prominent in the policy discourse stems from economics and the social sciences. Its driving concern is that AI may lead to technological unemployment, while there are also views holding that AI is well placed to augment workers'

capabilities and, thus, raise productivity and enable innovations. Accordingly, this literature focuses mainly on AI's potential to substitute for labour in the workplace and measures AI capabilities with regard to occupations and work tasks (the so-called task-based approach). Other strains of research from computer science and psychology analyse AI from the perspective of skills and abilities. They measure which computer capabilities are now available and how they will eventually relate to human skills.

Studies analysing AI capabilities may also differ with respect to the crudeness of their measures. Some measures are based on vague notions of the capabilities of computers. Others rely on ratings of whether AI can perform (more or less specifically defined) tasks. Yet others draw on actual AI performance.

The next few paragraphs provide a snapshot of the most prominent studies in each of these areas.

The task-based approach to measuring AI capabilities and its impact on jobs

Studies following the task-based approach analyse the extent to which AI can displace workers by focusing on occupations and their task content and by studying the susceptibility of tasks to automation. The goal is to determine the share of tasks within an occupation that can be performed by computers and, ultimately, the share of jobs in the economy that can be largely carried out by machines [see Frey and Osborne (2017[15]) or Brynjolfsson et al. (2018[16])]. In this literature, the focus is less on AI and more on its impact on work and employment. AI capabilities are rated with regard to broad descriptions of occupations or job tasks. These are usually derived from occupation taxonomies, such as the Occupational Information Network (O*NET) database of the US Department of Labor. The resulting measures of occupations' automatability are then matched to micro-level labour market data to further examine the characteristics of jobs at high risk of automation (e.g. industry, region and wage level), as well as the characteristics of workers who are at risk of displacement (e.g. age, gender and education level).

The origins of the task-based approach

The task-based approach has its origin in the seminal work of Autor, Levy and Murnane (2003[17]). Their study stipulates that machines can substitute for workers only in particular tasks. These are typically routine cognitive and manual tasks that follow exact repetitive, predictable procedures. As such, the tasks can be readily formalised and codified. By contrast, tasks that follow tacit, inexplicable rules, such as those involving flexibility, creativity, problem solving or complex interaction, are not apt for computerisation. Both types of tasks are operationalised loosely by using broad occupational descriptions. These include the involvement of direction, control and planning of activities at work or the extent to which finger dexterity is required.

Autor, Levy and Murnane (2003[17]) study how labour input in routine and non-routine tasks develops over time. The premise is that declining prices of technology should reduce labour demand for routine tasks and increase demand for non-routine tasks. That is, employers would increasingly replace workers in routine tasks with cheap machines. At the same time, they would employ workers for complementary non-routine tasks, such as developing, managing and monitoring machines.

The approach of Frey and Osborne

Since the Autor, Levy and Murnane study in 2003, computers have advanced significantly. They can now perform many of the tasks previously thought as "uncodifiable", such as driving or translation. This was possible mainly through progress in AI and in machine learning, in particular. To account for these technological advancements, Frey and Osborne (2017[15]) take a new perspective on measuring computer capabilities. Instead of asking what computers can do in the workplace, they concentrate on tasks that computers still cannot perform. As such tasks are declining in number, they are becoming easier to characterise and thus to assess. Specifically, the authors identify three engineering bottlenecks to AI-driven automation: perception and manipulation tasks, such as navigating in an unstructured

environment; creative intelligence tasks, such as composing music; and social intelligence tasks, such as negotiating and persuading. According to Frey and Osborne (2017[15]), computers can perform any task not subject to one of the three bottlenecks.

To quantify computers' ability to automate work, Frey and Osborne (2017[15]) first asked AI researchers to rate the automatability of 70 (of 700) occupations in the O*NET occupation taxonomy. O*NET contains systematic information on occupations' task content and skills requirements. Based on the task descriptions, the experts labelled occupations as automatable or non-automatable. In a second step, Frey and Osborne (2017[15]) approximated the three engineering bottlenecks with nine O*NET variables, available for the full set of occupations in the database. For example, the variable "originality", which describes the degree to which an occupation requires unusual or clever solutions, serves as a proxy for creative tasks. Finally, the authors estimated the relationship between the automatability of the initial 70 occupations (derived from the subjective expert assessments) and the bottlenecks (measured with the nine O*NET variables). They used the obtained estimates to predict the probability of automation of all 700 occupations.

The work of Frey and Osborne (2017[15]) stimulated much research but has several limitations. Most importantly, the information from O*NET used by experts to rate occupations' automatability involves simple one-line task descriptions. These descriptions provide little indication of task difficulty or specific examples. Based on the different tasks of an occupation, experts provided judgements for the entire occupation. This raises questions as to how they dealt with essential and less essential tasks of the job. Similarly, it is unclear whether they included abilities needed for the occupation but not listed in these databases because all (healthy) human beings have them (e.g. vision and common sense reasoning). In addition, such occupational-level judgements do not recognise that jobs within the same occupation may differ in their task mix and, hence, in their amenability to automation. Furthermore, only occupations where experts were most confident were considered; only occupations in which all tasks were rated automatable were labelled as automatable. However, the study does not specify how inter-rater agreement was determined, and how high it was. Neither does the work address the issue of those occupations that are not fully automatable but have a high number of automatable tasks. The lack of such information raises questions about the validity of the exercise.

Two studies supported by the OECD – Arntz, Gregory and Zierahn (2016[5]) and Nedelkoska and Quintini (2018[6]) – address one of these points: they estimate the risk of automation at the level of jobs instead of occupations. More precisely, the studies map the expert judgements on the automatability of the 70 occupations from the study of Frey and Osborne (2017[15]) to micro-level data from the Survey of Adult Skills of OECD's Programme for the International Assessment of Adult Competencies (PIAAC). They then estimate the link between automatability of occupations and various work tasks at the level of individual jobs assessed in PIAAC. This contrasts with Frey and Osborne (2017[15]), who model occupations' automatability as a function of hard-to-automate, bottleneck tasks at the occupation level. Analysing how job-level characteristics are linked to automation makes it possible to reflect the variation of jobs within occupations in the overall estimate of automatability across the economy.

Further efforts to assess the automation of tasks

Following the task-based approach, some major consultancies have issued reports on the impact of technology on employment and the economy. McKinsey Global Institute studies automation by linking 2 000 work activities to 18 key performance capabilities needed to execute them (such as sensory perception and retrieving information) and by establishing the level of performance that AI has with regard to these capabilities (Manyika et al., 2017[18]). However, the study does not describe the methodological approach in further detail. It remains unclear how capabilities are defined and matched to abilities. Similarly, it is unknown how their susceptibility to automation is determined. In addition, PricewaterhouseCoopers (2018[19]) examines the global economic impact of AI by studying both its

potential to enhance productivity through automation and to improve product quality. To assess how AI can automate jobs, the report adopts the approach of Arntz, Gregory and Zierahn (2016[5]) and PwC (PwC, 2017[20]).

Other studies in this literature focus more narrowly on AI. Brynjolfsson, Mitchell and Rock (2018[16]), for example, define key criteria for whether a work task is suitable for machine learning applications. These criteria include, for example, clearly definable rules and goals or the availability of large digital datasets for training algorithms (see also Brynjolfsson and Mitchell (2017[21])). The authors then rate 2 069 work activities linked to 964 occupations in O*NET against these criteria to measure occupations' suitability for machine learning. By contrast, Felten, Raj and Seamans (2019[22]) use objective AI metrics made available by the Electronic Frontier Foundation to track progress of AI across major application domains, such as image and speech recognition. They map these AI progress measures to information from O*NET on key abilities required in occupations. To that end, they ask gig workers on a freelancing platform to rate the relatedness between both. In this way, the authors assess the extent to which occupations are exposed to computerisation.

While all these studies rely in some way on the subjective judgement of computer experts, economists or "laypersons", the study of Webb (2020[23]) aims at developing an objective measure of the applicability of technology in the workplace. To derive such a measure for AI, the study first identifies patents of AI technologies by scanning patent descriptions for keywords such as "neural networks" and "deep learning". It compares verb-noun pairs in patents' texts and occupations' task descriptions available in O*NET. In this way, it quantifies the overlap between such AI patents and occupations.

Squicciarini and Staccioli (forthcoming[24]) adopt a similar approach to Webb (2020[23]) but with a focus on robotics. They identify patents associated to labour-saving robotic technologies by applying text-mining techniques. Subsequently, they connect these patents to occupation descriptions available in ISCO08 – a standardised classification of occupations. They use a text similarity algorithm to measure occupations' exposure to such innovations.

Again, these studies rely on broad occupation descriptions available in occupation taxonomies. They focus more on quantifying the extent to which AI can automate the economy than on developing a comprehensive measurement of what computers can do.

Remaining gaps in methodology

To sum up, research following the task-based approach has so far mostly focused on how evolving AI and robotics alter the workplace by automating tasks within occupations. Most studies following this approach share some (or all) of the following methodological gaps:

- judgements are based on vague descriptions of skills or tasks, which omit important details needed to evaluate if AI can do them
- judgements about entire jobs mix tasks that AI can and cannot do
- judgements about entire jobs require knowledge of AI capabilities and knowledge of job design at the same time, but no experts have both
- information about the experts' identity, selection and domains of expertise is lacking
- information about the exact methodology and the rating process is lacking.

In addition, the task-based approach offers a narrow view of humans: workers are seen as displaced from tasks or from entire jobs that have been overtaken by computers. This leaves a number of key questions largely unexplored. How do people's skills and abilities compare to AI performance? Which of these skills are reproducible? Which can be usefully complemented or augmented by machines? Which skills are hard to automate and, thus, worth investing in? Skills-based approaches in computer science and psychology offer a promising way forward to address these gaps. This new approach is discussed next.

Skills-based assessment: A new approach

Apart from the work in economics that assesses AI capabilities as an initial step to studying its potential impact on jobs, there are efforts carried out to assess those capabilities in the field of computer science (Martínez-Plumed et al., 2021[25]; Clark and Etzioni, 2016[26]; Crosby, Beyret and Halina, 2019[27]; Ohlsson et al., 2016[28]; Davis, 2016[29]). In addition, a wide set of assessment approaches has been developed over the past century to measure human capabilities that could potentially be applied to understanding AI. These tools offer a deeper way of measuring and understanding AI capabilities that goes beyond judgements about the automatability of entire occupations in the economics literature.

In 2016, an OECD/CERI pilot project assessed computers' capabilities using an assessment of human competences linked to the workforce. This was an initial exploration of how to connect a more skills-based assessment of AI capabilities to the kinds of economic questions that concern policy makers. The pilot assessed computers' literacy, numeracy and problem-solving skills using the PIAAC test. This test assesses skills in the adult population and is linked to information about work and other adult activities (Elliott, 2017[30]). It is part of the OECD's effort to evaluate educational outcomes through assessing skills such as literacy and numeracy. Some of these assessments, including both PIAAC and the Programme for International Student Assessment have been used for a long time at a large scale across different contexts. As such, they provide robust information on the distribution of these skills in the population. Assessing computer capabilities through such tests potentially allows for comparing AI and robotics skills with humans at various proficiency levels.

To gauge the changes that technology may bring, the pilot study asked a group of computer scientists to rate the ability of current computer technology to answer each test question. In addition, a subset of experts also provided projections for the evolution of technology in the next ten years. The aggregate rating across specific test questions was then used to place AI capabilities on the PIAAC test scale. This helped understand how AI capabilities compare to the human population in these three skill areas.

Strengths of the new approach

The pilot study revealed a number of strengths of this new approach.

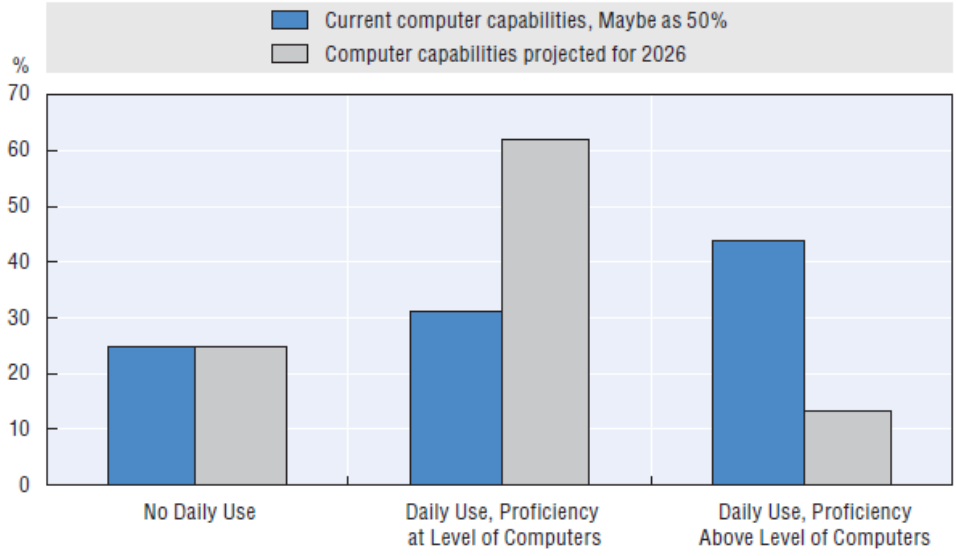
First, *rating with regard to specific test items provides a more precise estimation* of computer capabilities. As discussed in the section above, prior estimations related to general descriptions of work tasks or activities, or even broader descriptions of occupations. Experts judging computer capabilities do not know exactly what granular tasks are required to carry these out. Their judgements thus inevitably involve assumptions that are likely to vary greatly across experts (Elliott, 2017[30]). By contrast, questions in standardised tests are precise and contextualised. This allows computer scientists to analyse the information processing required to answer a specific question based on the information provided (Elliott, 2017[30]). This level of specificity implies greater reliability across raters and greater reproducibility.

Second, *using human tests makes it possible to compare computer and human capabilities*. In particular, when large-scale data on human skills are available across different contexts, different age groups and occupations, these data can be used to conduct fine-grained analyses of skill supply and future demand. Simply put, employers could rearrange job tasks to use AI for its capabilities and human workers for the capabilities that AI still lacks. This means that skill demand for human workers should shift towards the (aspects of) capabilities that AI still lacks. A fine-grained comparison of human and computer capabilities across the full range of skills required for work and life will help avoid jumping to faulty conclusions on job automation. Instead, this connection provides information about AI's impacts that extends beyond the definition of current occupations. It will be useful for thinking about AI's implications for employment more generally, including occupations that do not yet exist, as well as education.

The pilot study illustrates this potential through a number of figures and analyses. PIAAC questions are grouped in five levels based on their difficulty for the adult human population. Expert ratings can indicate

whether computers can perform at a certain level of difficulty as human adults. Figure 1.1, for example, shows the distribution of workers based on whether they use general cognitive skills daily in their work and, if they do, how they compare to computer capabilities. Work tasks of the part of the workforce that does not use any of these skills on a daily basis will not be substantially affected by the computer capabilities examined in the pilot study (as the first two bars in the figure show). Workers who use one or more of these skills regularly and have proficiency above the projected level of computer capabilities will likely continue to have regular tasks using these skills that are not substantially affected by computer capabilities in these areas (as the last two bars in the figure show). However, automation will likely affect those who use one or more of these skills on a daily basis but have proficiencies only at the level of projected computer capabilities (middle bars) (Elliott, 2017[30]).

Figure 1.1. Distribution of workers by use of general cognitive skills and proficiency compared to computers (Results from the pilot study)



Source: (Elliott, 2017, p. 92[30]).

Challenges and further development of the new approach

The pilot study also identified a number of challenges of the approach. Overfitting is a commonly cited danger of assessing whether computer technology can answer a particular test item. It is often possible to train computers for a specific task. However, this does not mean the computer can perform a range of similar tasks (Elliott, 2017[30]). Overfitting relates to the question of generalisability, i.e. the possibility to infer that computers have an underlying capability from their performance on specific items. Another challenge is whether the same ability is tested through a test item for computers and for humans. For example, an item that requires counting objects in a picture tests a simple numeracy skill for humans. However, the challenge for AI is visual processing rather than counting (Elliott, 2017[30]).

To sum up, despite the challenges identified in the pilot study, the new approach for assessing AI capabilities is promising both in terms of its credibility (validity and reliability) and in producing measures that are meaningful for policy. Further developing and extending this approach could enlarge the conversation about AI capabilities. On the one hand, it could estimate automatability indices for current occupations (which has been done before). However, it would also provide information on how occupations are likely to transform, what new occupations may emerge and what this all means for developing people's

skills within and outside formal education. The following section discusses what this development and extension involve.

Purpose and structure of the report

The pilot work explored assessing AI capabilities with one example test that involves just a few skills. As a result of its preliminary success, CERI decided to expand the work to a more comprehensive set of assessments. However, the methodology in the pilot study needs to be refined and the assessment extended to a comprehensive list of capabilities. This is necessary to establish valid, reliable and meaningful measures for an ongoing systematic assessment of machine capabilities. This work involves two steps:

- reviewing taxonomies of human skills and capabilities[1], and identifying an appropriate taxonomy to use for the project that spans the full range of skills used in the workplace
- reviewing available tests of human skills and establishing criteria for their suitability for assessing AI capabilities.

The project needs a comprehensive framework of skills that would fulfil three requirements. First, it would include all the skills that people need for their work and life. Second, it would be suitable for analysing AI capabilities, including both those similar to and different from human skills. Third, it would be suitable for comparing these capabilities to human skills and draw implications of AI progress on the world of work and education.

As demonstrated by the challenges described earlier, assessing humans and machines is a different exercise. Therefore, the project needs to bring together different disciplines and interpret their findings for one another. This technical report is a first step in this process.

The report builds on an online meeting of the AIFS project held on 5-6 October 2020 with experts from various domains of psychology and computer science. The meeting sought to explore different domains of psychology (cognitive, personality, industrial and occupational, developmental and neuropsychology) and to review existing taxonomies of human skills in these domains. The meeting also aimed to identify tests of these skills and discuss their strengths, weaknesses and applicability for assessing machine capabilities.

Psychologists were asked to present a taxonomy (or taxonomies) of the skill domain of their expertise and describe the types of tests available to assess these skills. Experts were requested to discuss the challenges and opportunities of assessing these skills based on the research literature. The experts then provided sufficient examples of actual test questions to illustrate the kinds of tasks typically included and the criteria used to evaluate performance. Computer scientists were invited to reflect on the progress of AI and robotics technologies. They presented types of empirical evaluations and benchmarks in the field, and outlined the main considerations for assessing AI and robotics capabilities against human capabilities. Experts then compared the different types of skill taxonomies and tests with the objective to work towards some broadly supported guidelines to govern the project's choice of a skill taxonomy and a set of tests.

This volume contains papers prepared by psychologists and computer scientists who attended the meeting. They each present research from their specific field of expertise and reflect on the project based on the exchange in the meeting and their personal-professional views. Content sometimes overlaps between the chapters. There is also some repetition both in the arguments and in how these are illustrated. Views across some chapters may also either complement or conflict with each other. Such recurrence, complementarity and conflict of arguments are an invaluable resource for the project and are necessary to elicit guidelines for the way forward.

Report structure

Part I. Setting the scene

The first part sets the scene for the report with two introductory chapters.

The current **Chapter 1** presents the background and rationale for this work.

In **Chapter 2**, *Kenneth Forbus* analyses the progress in AI over the past four decades. The author describes three ongoing revolutions – deep learning, knowledge graphs and reasoning – and foresees a fourth revolution: that of integrated intelligence. The chapter discusses the implications of these revolutions for efforts to derive relevant measures of AI's progress with respect to human capabilities.

Part II. Taxonomies and tests of human skills

The second part explores taxonomies of human skills in different branches of psychology and reviews existing measures of these skills. It distinguishes two major areas: cognitive psychology discussed in Chapters 3 to 7 and industrial-organisational psychology addressed in Chapters 8 to 10.

Cognitive abilities and their extensions

Chapter 3 by *Patrick Kyllonen* provides an overview of widely known taxonomies of human cognitive abilities, presents the history of their development, and discusses their strengths and weaknesses. The author describes measures of cognitive abilities and their quality characteristics such as reliability, validity, fairness and measurement invariance. Finally, the chapter discusses the prospects and feasibility of using these tests as the basis for evaluating machine intelligence.

In **Chapter 4**, *Sylvie Chokron* presents the neuropsychological perspective of capturing weaknesses and strengths of cognitive abilities in children. The author describes the most commonly used neuropsychological tests, as well as their limits and caveats in understanding the cognitive profile of children. The chapter also considers the opportunities and challenges in using such tests for assessing machine capabilities.

Chapter 5 focuses on social and emotional skills. *Filip De Fruyt* reviews the theoretical conceptualisations of social and emotional skills, and discusses how these relate to educational and labour-market outcomes. The paper presents taxonomies of these skills, including the widely known Big Five framework, and describes different types of items through which such skills can be assessed. Importantly, the author discusses recent developments in the field. These attempt to provide more objective measures of social and emotional skills, such as situational judgement tests and behavioural residue indicators.

Chapter 6 by *Anita Woolley* explores the recent concept of collective intelligence, i.e. the ability of a group to perform a wide range of tasks. The paper presents task batteries to measure collective intelligence and describes how these can be used to elicit the factors that predict team performance. Ongoing research also includes understanding how AI capabilities can enhance collective intelligence in a mixed team of machines and humans. Finally, the author illustrates how these measures can provide a vehicle for assessing artificial social intelligence.

In **Chapter 7**, *Samuel Greiff* and *Jan Dörendhal* describe two major skill domains typically measured in large-scale educational assessments: core domain skills such as mathematics, reading and science literacy, and transversal skills such as problem solving, collaboration and creativity. The chapter presents the theoretical underpinning and measurement of these skills and examines their role in occupational settings. The chapter concludes with recommendations regarding the use of education tests for assessing AI capabilities.

Occupational assessments

Chapter 8 proposes an assessment strategy that draws on tests developed for jobs subject to licensing examination. *Phillip Ackerman* presents the foundations of human intelligence tests. The author discusses a number of methodological challenges including those arising from tacit knowledge, humans' use of tools, differences in learning between humans and AI, and the inaccuracy of skills assessments at high performance levels. Based on these challenges, the chapter argues that the OECD project should focus on domain knowledge and skills – in the context of specific jobs – rather than higher-order cognitive abilities. .

Chapter 9 provides a vocational perspective to skills assessment. *Britta Rüschoff* reviews the methods of skills assessment in German vocational education and training (VET). The chapter defines vocational competences and presents instruments to assess them in VET examinations through concrete examples. The author also describes how these examinations are developed and administered, and discusses the validity and reliability of the instruments. Finally, the chapter indicates the advantages of using VET tests for assessing AI capabilities. It concludes with considerations for applying these instruments to machines.

In **Chapter 10**, *David Dorsey* and *Scott Oppler* propose an approach for comparing human and AI capabilities based on comprehensive occupational taxonomies. The authors start with clarifying the structure and underlying concepts of occupational databases. The chapter then outlines a number of methodological recommendations and describes four major steps of the proposed approach: identifying an occupational taxonomy, sampling occupations from the taxonomy, collecting expert judgement on AI capabilities and analysing data from expert interviews.

Part III. AI capabilities and their measures

The third part of the volume explores the perspective of computer scientists on the evaluation of AI and robotics capabilities. Chapters 11 to 13 discuss major challenges in assessing AI capabilities with human tests, while Chapters 14 to 17 focus on existing empirical evaluation efforts of machines.

Challenges of assessing AI capabilities with human tests

Chapter 11 connects the second and the third part of the report. After situating AI measurement in the context of the roles AI can play in the future, *José Hernández-Orallo* provides an overview of human skill taxonomies and links these to the world of AI. The chapter explores types of human tests used in recruitment and education, and contrasts these with the evaluation of machines. The author discusses the challenges of using human tests for assessing AI capabilities and identifies guidelines for devising tests that can compare the capabilities of humans and AI reliably.

Chapter 12 focuses on the specificities of machines and their striking differences from humans. *Ernest Davis* presents areas in which computers excel and illustrates their weaknesses compared to humans through examples of sometimes "grotesque" failures. The author proposes consequences of these various strengths and limitations for using human tests for assessing AI.

Chapter 13 also brings attention to the ways in which AI is similar to and utterly different from human intelligence. *Richard Granger* first discusses how the architectures of artificial neural networks relate to networks of neurons in the human brain. The author then compares the behaviour and computational abilities arising from artificial neural networks to human behaviour and abilities, illustrating both with entertaining examples. The chapter points to current efforts to overcome shortcomings and concludes with a number of implications for understanding machine capabilities.

Efforts to assess AI and robotics capabilities

In **Chapter 14**, *Anthony Cohn* presents approaches and methods of the AI community to measuring and evaluating AI systems. The author first presents tests proposed for measuring AI, then describes some competitions created to compare AI systems, as well as a few benchmark datasets. The chapter discusses some of the benefits and limitations of these approaches.

Chapter 15 focuses on empirical evaluations of AI systems as performed by the French Institute of Metrology (Laboratoire national de métrologie et d'essais: LNE). *Guillaume Avrin* first presents the characteristics and process of these evaluations. The author then proposes a high-level taxonomy of AI capabilities and generalises it to other AI tasks to draw a parallel with human capabilities. The chapter then discusses the relevance of existing evaluation methods for comparing AI and human capabilities. It concludes with recommendations for the project approach.

Chapter 16 provides an overview of evaluation techniques applied in the domain of natural language processing. *Yvette Graham* describes methods that offer fair and replicable evaluations of system performance in this domain. The author shows how longitudinal evaluation can capture progress in AI language processing capabilities and how these methods allow for comparison with human performance. The chapter also discusses human-machine hybridisation in tasks and its implication for understanding the potential for machines in society.

In **Chapter 17**, *Lucy Cheke*, *Marta Halina* and *Matthew Crosby* focus on basic or common sense skills that all healthy human adults have but in which machines still often fail. The authors propose a taxonomy of these skills identifying two major domains: spatial and social skills, and describe tests used in the fields of animal and developmental psychology. The chapter presents examples of workplaces and situations that might require the use of such skills, and explore limitations and opportunities for assessing common sense skills in AI.

Part IV. Reflections and a pragmatic way forward

The last part of the report attempts to synthesise the discussion and draw conclusions as to the way forward for the AIFS project.

In **Chapter 18,** *Art Graesser* discusses three questions relevant to the AIFS project: What is the value in identifying ideal models when comparing humans and AI and robotic systems? How might we conduct a systematic mapping between skill taxonomies, tasks, tests and functional AI components? How can we handle major differences in the skills we target, the different occupations and changes in the worlds we live in? The author offers suggestions on next steps in addressing these questions.

Chapter 19 provides guidance for setting up a general analytical framework for assessing AI capabilities with regard to human skills. *Eva Baker* and *Harry O'Neil* offer recommendations on operative aspects of the project, on the selection of tests for comparing AI and human skills, and on the selection and training of expert raters. The chapter concludes with a summary of considerations made for planning the study.

Finally, in **Chapter 20**, *Stuart Elliott* reflects on key considerations from the expert contributions in the field of psychology and computer science. The author proposes to bring together the different domains of psychology to benefit from the strength and relevance of each domain. The chapter also suggests a pragmatic way forward for the project to address the concerns formulated by the AI community and develop AI-specific assessment approaches where those are required.

References

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*, No. 189, OECD Publishing, Paris, https://dx.doi.org/10.1787/5jlz9h56dvq7-en. [5]

Autor, D., F. Levy and R. Murnane (2003), "The skill content of recent technological change: An empirical exploration", *The Quarterly Journal of Economics*, Vol. 118/4, pp. 1279-1333, http://dx.doi.org/10.1162/003355303322552801. [17]

Baruffaldi, S. et al. (2020), "Identifying and measuring developments in artificial intelligence: Making the impossible possible", *OECD Science, Technology and Industry Working Papers*, No. 2020/05, OECD Publishing, Paris, https://dx.doi.org/10.1787/5f65ff7e-en. [3]

Bresnahan, T. and M. Trajtenberg (1992), *General Purpose Technologies "Engines of Growth?"*, National Bureau of Economic Research, Cambridge, MA, http://dx.doi.org/10.3386/w4148. [12]

Brynjolfsson, E. and T. Mitchell (2017), "What can machine learning do? Workforce implications", *Science*, Vol. 358/6370, pp. 1530-1534, http://dx.doi.org/10.1126/science.aap8062. [21]

Brynjolfsson, E., T. Mitchell and D. Rock (2018), "What can machines learn and what does It mean for occupations and the economy?", *AEA Papers and Proceedings*, Vol. 108, pp. 43-47, http://dx.doi.org/10.1257/pandp.20181019. [16]

Clark, P. and O. Etzioni (2016), "My Computer Is an Honor Student — but How Intelligent Is It? Standardized Tests as a Measure of AI", *AI Magazine*, Vol. 37/1, pp. 5-12, http://dx.doi.org/10.1609/aimag.v37i1.2636. [26]

Crosby, M., B. Beyret and M. Halina (2019), "The Animal-AI Olympics", *Nature Machine Intelligence*, Vol. 1/5, pp. 257-257, http://dx.doi.org/10.1038/s42256-019-0050-3. [27]

Davis, E. (2016), "How to Write Science Questions that Are Easy for People and Hard for Computers", *AI Magazine*, Vol. 37/1, pp. 13-22, http://dx.doi.org/10.1609/aimag.v37i1.2637. [29]

Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264284395-en. [30]

Felten, E., M. Raj and R. Seamans (2019), "The variable impact of artificial intelligence on labor: The role of complementary skills and technologies", *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.3368605. [22]

Frey, C. and M. Osborne (2017), "The future of employment: How susceptible are jobs to computerisation?", *Technological Forecasting and Social Change*, Vol. 114, pp. 254-280, http://dx.doi.org/10.1016/j.techfore.2016.08.019. [15]

Lane, M. and A. Saint-Martin (2021), "The impact of Artificial Intelligence on the labour market: What do we know so far?", *OECD Social, Employment and Migration Working Papers*, No. 256, OECD Publishing, Paris, https://dx.doi.org/10.1787/7c895724-en. [7]

Manyika, J. et al. (2017), *A Future That Works: Automation, Employment, and Productivity*, McKinsey Global Institute, New York. [18]

Martínez-Plumed, F. et al. (2021), "Research community dynamics behind popular AI benchmarks", *Nature Machine Intelligence*, Vol. 3/7, pp. 581-589, http://dx.doi.org/10.1038/s42256-021-00339-6. [25]

Nakazato, S. and M. Squicciarini (2021), "Artificial intelligence companies, goods and services: A trademark-based analysis", *OECD Science, Technology and Industry Working Papers*, No. 2021/06, OECD Publishing, Paris, https://dx.doi.org/10.1787/2db2d7f4-en. [8]

Nedelkoska, L. and G. Quintini (2018), "Automation, skills use and training", *OECD Social, Employment and Migration Working Papers*, No. 202, OECD Publishing, Paris, https://dx.doi.org/10.1787/2e2f4eea-en. [6]

OECD (2021), "Why accelerate the development and deployment of robots?", in *OECD Science, Technology and Innovation Outlook 2021: Times of Crisis and Opportunity*, OECD Publishing, Paris, https://dx.doi.org/10.1787/0901069e-en. [14]

OECD (2020), *The OECD Artificial Intelligence Policy Observatory - OECD.AI*, https://oecd.ai/ (accessed on 20 January 2021). [2]

OECD (2019), *Artificial Intelligence in Society*, OECD Publishing, Paris, https://dx.doi.org/10.1787/eedfee77-en. [4]

OECD (2019), *Recommendation of the Council on Artificial Intelligence*, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#supportDocuments (accessed on 20 January 2021). [1]

OECD (2017), *OECD Digital Economy Outlook 2017*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264276284-en. [13]

Ohlsson, S. et al. (2016), "Measuring an artificial intelligence system's performance on a Verbal IQ test for young children", *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 29/4, pp. 679-693, http://dx.doi.org/10.1080/0952813x.2016.1213060. [28]

PwC (2018), *The Macroeconomic Impact of Artificial Intelligence*, PricewaterhouseCoopers, London, https://www.pwc.co.uk/economic-services/assets/macroeconomic-impact-of-ai-technical-report-feb-18.pdf. [19]

PwC (2017), *Will Robots Really Steal our Jobs? The Potential Impact of Automation on the UK and other Major Economies*, PricewaterhouseCoopers, London. [20]

Squicciarini, M. and J. Staccioli (forthcoming), "Labour-saving technologies and employment levels: Are robots really making workers redundant?", *OECD Science, Technology and Industry Working Papers*. [24]

van der Vlies, R. (2020), "Digital strategies in education across OECD countries: Exploring education policies on digital technologies", *OECD Education Working Papers*, No. 226, OECD Publishing, Paris, https://dx.doi.org/10.1787/33dd4c26-en. [10]

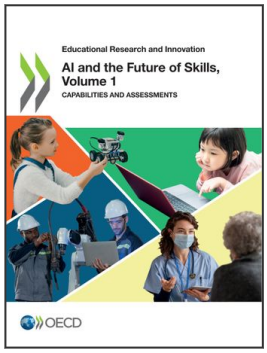
Verhagen, A. (forthcoming), "Opportunities and drawbacks of using Artificial Intelligence for training", *OECD Social, Employment and Migration Working Papers*. [11]

Vincent-Lancrin, S. and R. van der Vlies (2020), "Trustworthy artificial intelligence (AI) in education: Promises and challenges", *OECD Education Working Papers*, No. 218, OECD Publishing, Paris, https://dx.doi.org/10.1787/a6c90fa9-en. [9]

Webb, M. (2020), "The impact of artificial intelligence on the labor market", *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.3482150. [23]

Notes

[1] This report will not distinguish between the terms "skills", "abilities", "capabilities" and "competences". They will be used interchangeably. Single chapters, however, may adopt more precise definitions, which will be made explicit.



From:

AI and the Future of Skills, Volume 1

Capabilities and Assessments

Access the complete publication at:

https://doi.org/10.1787/5ee71f34-en

Please cite this chapter as:

Révai, Nóra, Mila Staneva and Abel Baret (2021), "New approaches to understanding the impact of computers on work and education", in OECD, *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/65774a32-en

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Extracts from publications may be subject to additional disclaimers, which are set out in the complete version of the publication, available at the link provided.

The use of this work, whether digital or print, is governed by the Terms and Conditions to be found at http://www.oecd.org/termsandconditions.