1 Overview

Mila Staneva, OECD

The AI and the Future of Skills (AIFS) project at OECD's Centre for Education Research and Innovation (CERI) aims at developing a comprehensive and authoritative approach to regularly measuring artificial intelligence (AI) capabilities and comparing them to human skills. This chapter provides an overview of the project, outlining its goals, past activities and future directions. It describes the second stage of AIFS (2021-22), which is the subject of this volume. This stage explored three sources of information for assessing AI capabilities: collecting expert judgement on AI performance on education tests, collecting experts' evaluations of AI on complex occupational tasks and using existing measures from direct evaluations of AI systems.

Artificial intelligence (AI) and robotics[1] are evolving rapidly, propelled by steady innovative breakthroughs. The result is an ever-expanding scope of applications, covering domains as varied as health care, finance, transportation and education. More recently, the introduction of ChatGPT, a sophisticated AI chatbot, provided a quintessential illustration of this rapid advancement. ChatGPT's remarkably human-like interactions and contextual sensitivity underscore the considerable strides achieved in AI just over a short period of time. Its ability to perform a variety of tasks, such as answering questions, composing poetry and music, and writing and debugging code, illustrates its wide application. This has triggered debates over the potential impact of AI on the economy and society, both in research and policy spheres, as well as in the media.

Understanding how AI can affect the economy and society – and the education system that prepares students for both – requires an understanding of the capabilities of this technology and their development trajectory. Moreover, AI capabilities need to be compared to human skills to understand where AI can replace humans and where it can complement them. This knowledge base will help predict which tasks AI may automate and, consequently, how AI may shift the demand for skills and challenge employment and education. Policy makers can use this information to reshape education systems in accordance with future skills needs and to develop tailored labour-market policies.

The AI and the Future of Skills (AIFS) project at OECD's Centre for Education Research and Innovation (CERI) is developing a comprehensive and authoritative approach to regularly measuring AI capabilities and comparing them to human skills. The capability measures will cover skills important in the workplace and everyday life, and developed in education systems. Ideally, they will provide a common ground for policy discussions about the potential effects of AI by establishing an accepted and accessible framework to describe AI capabilities and their change over time.

The first stage of AIFS explored ways to categorise AI capabilities and existing tests to assess them. The project reviewed numerous skill taxonomies and skill assessments from the fields of cognitive psychology, industrial-organisational psychology, animal cognition, child development, neuropsychology and education. In addition, it considered AI evaluations developed and used in computer science. To that end, the project identified and interviewed key experts from multiple disciplines to ensure the developed methodology includes all relevant perspectives and expertise domains. These experts explored the usefulness of existing taxonomies and tests for assessing the capabilities of AI and robotics and comparing them to human skills. The results of this work are presented in the project's first methodological report (OECD, 2021[1]).

The present report describes the second stage of developing the methodology of the AI assessment. In this stage, the project conducted exploratory assessments of AI in three domains identified as key in the preceding phase. The project started by exploring methods for eliciting expert knowledge on AI capabilities. First, it collected expert judgement on whether AI can solve education tests developed for humans. Education tests provide a useful way to compare AI to human capabilities in domains relevant to education and work. Second, the project asked experts to evaluate AI on complex occupational tasks. These tasks stem from tests used to certify workers for occupations and provide insights in AI's readiness for real-world applications. Third, the project moved to exploring the use of measures from direct evaluations of AI systems developed in computer research. These direct measures are more objective than ones relying on expert judgements but do not cover the full spectrum of skills relevant in work and education.

The three exploratory efforts were carried out separately from each other. In the next project stage, these strands of work will be integrated into developing measures of AI capabilities. These measures will quantify the current state-of-the-art of AI technology with regard to several key capabilities. The plan is to regularly update them to track progress in AI and gradually expand them to cover new capability domains. Importantly, the measures will be linked to existing occupational and skill taxonomies to enable analyses of the implications of evolving AI for work and skills development.

This chapter introduces the AIFS project, including its goals, past activities and future directions. It then recapitulates results from the initial stage of the project and shows how this work evolved in the second stage, the focus of this report. The chapter describes the three exploratory efforts carried out at this stage in further depth. It concludes with an outline of the structure of the report.

Overview of the AI and the Future of Skills project

Project goals

AIFS is premised on the idea that policy makers and the public can benefit from measures of AI capabilities that are comparable to macroeconomic indicators, such as gross domestic product growth, price inflation or unemployment rate. Like the latter, AI measures should provide a high-level understanding of complex developments related to AI to non-experts. They should support decisions on whether and what policy interventions may be needed as further substantial changes in AI take place.

As with any measures, the AI capability measures should be valid, reliable and fair. In other words, they should reflect the capabilities of AI they claim to measure (validity), provide consistent information (reliability) and consider different AI systems equally (fair). Beyond these general measurement qualities, measures aiming at informing policy makers and the public on AI should meet several additional criteria:

- ***Understandable***

AI measures should be easy to interpret. They should signal strengths and limitations of AI in a straightforward manner, understandable to non-experts. This requirement suggests a small set of measures, 5 to 10, that condense a wealth of information on AI trends. The scales of these measures should convey meaningful contrasts in performance. They should be summarised into a small number of performance levels that include qualitative descriptions of what AI can do at the respective level.

- ***Comprehensive***

The measures should cover all key aspects of AI needed for understanding its likely large-scale implications. This requirement does not contradict the goal of reducing complexity by providing only a small number of AI measures. The measures will be constructed out of many components, which could be used on their own to provide a more detailed picture to interested users. The choice of the components and the way they will be aggregated into final measures will be guided by a carefully developed conceptual framework.

- ***Repeatable***

The measures need to indicate change in AI, which calls for repetition at regular intervals. This is important because AI is changing quickly, and decision makers need to be informed when major surges in technology occur. This requirement means the assessment must be feasible to reproduce. That is, the assessment instruments must be standardised and reliable. The assessment itself must be institutionally embedded and supported by an established process for receiving input from experts.

- ***Policy relevant***

The measures should enable conclusions about AI's potential impact on education, employment and the economy. This requires that AI measures compare AI and human capabilities. This comparison would show how AI is likely to change the role of humans in carrying out different tasks (e.g. by replacing them or by providing extensive support that transforms the human role and its skills requirements). This would help policy makers understand AI's implications for work, education and society.

Past and current activities

The AIFS project was preceded by an OECD pilot study in 2016 (Elliott, 2017[2]). The study collected expert judgement on whether AI can carry out education tests designed for humans. It used OECD's Survey of Adult Skills, which is part of the Programme for International Assessment of Adult Competencies (PIAAC). PIAAC tests adults' proficiency with respect to three core skills – literacy, numeracy and problem solving.[2] The pilot study served as a stepping stone into the AIFS project, setting the focus on assessing AI in key skill domains of humans using expert evaluations.

In 2019-20, the AIFS project started by reviewing existing skill taxonomies and the tests developed to assess them. The goal was to expand the approach set out in the pilot study into a comprehensive AI assessment across the whole range of skills relevant for work and education. The results of this work are presented in project's first methodological report (OECD, 2021[1]). The volume contains 18 chapters by experts from various domains of computer science and psychology, offering perspectives on capability taxonomies and assessments used in their fields. This work shifted the focus of the project to relying more heavily on measures developed in AI research that are based on direct evaluations of AI systems.

In 2021-22, the AIFS project tested assessment approaches identified as key in the preceding project phase. This work – the subject of the current report – consists of several exploratory studies in three domains. First, the project continued to explore methods for collecting expert judgement about AI performance on education tests. Second, it expanded this assessment on complex occupational tasks from occupation entry examinations. Third, it explored the use of measures derived from direct assessments of AI systems. These exploratory efforts involved a series of expert meetings and expert surveys:

- Expert knowledge elicitation (March 2021): expert meeting to discuss the challenges and solutions of gathering direct measures on AI and robotics capabilities using human tests.
- Direct measures of AI capabilities (July and October 2021): expert meetings and commissioned work to explore ways for selecting and systematising existing direct measures of AI capabilities in the field.
- Follow-up of the pilot study with PIAAC (December 2021): an expert survey and workshop to collect expert judgement on AI capabilities in literacy and numeracy.
- Framing the rating exercise for experts (March 2022): an expert meeting to discuss a revised approach to instructing experts to rate potential AI performance on human tests.
- Second round of the follow-up study with PIAAC (September 2022): an expert survey and workshop to collect expert ratings on AI performance in numeracy using a revised framing of the rating exercise.
- Study using Programme for International Student Assessment (PISA) tests (June 2022): a large-scale survey to collect expert ratings on AI performance in science using a revised approach for expert knowledge elicitation.
- Occupational tasks (July and September 2022): two expert meetings to discuss possible approaches to providing expert judgement on AI on a set of occupational performance tasks.

The third stage of the project, 2023-24, is integrating the three strands of exploratory work into a coherent approach for assessing AI capabilities. It is developing several measures of key AI capabilities that will be linked to occupational taxonomies and taxonomies of human skills. In addition, the project is developing two in-depth studies of AI implications for work and education. The first study will focus on a few exemplary work tasks to examine how they can be redesigned to enable human-AI collaborations. The second study will look at the ways evolving AI can support and transform the capabilities developed in formal education.

The subsequent sections describe the lessons learnt during the first stage of the project and how they evolved into the three exploratory efforts that are the subject of this report.

Lessons learnt from the first project stage

The first stage of the project aimed to identify AI capabilities to be assessed, as well as tests that could be used to assess them (OECD, 2021[1]). Experts from a variety of disciplines were invited to review and propose resources for this purpose. The result was a conceptual framework that summarises the available skill taxonomies and assessments into three major types (see Figure 1.1).

Figure 1.1. Sources of AI assessments

Source: Elliott, S. (2021[3]), *"Building an assessment of artificial intelligence capabilities"*, in AI and the Future of Skills, Volume 1 https://doi.org/10.1787/01421d08-en.

First, experts discussed taxonomies and tests developed to assess isolated human skills (bottom left in Figure 1.1). The pilot work that preceded the AIFS project has explored such resources by collecting expert judgement on AI capabilities in literacy, numeracy and problem solving using an OECD education test (Elliott, 2017[2]). Next to skills assessments in education, experts reviewed work from psychology related to assessing numerous other skills, such as socio-emotional, psychomotor or perceptual skills. In addition, tests from the fields of animal cognition and child development were proposed for assessing AI in basic low-level skills that all healthy adult humans share (e.g. spatial and episodic memory).

Human tests are a promising tool for assessing AI in many regards. They are standardised, objective and repeatable, and allow for comparisons of AI to human performance in key skill domains. However, experts expressed concern that these tests are not explicitly designed for machines. Consequently, they may omit important characteristics of AI performance. Moreover, the psychometric assumptions upon which they rely do not necessarily hold for machines. That is, high performance of AI on one task does not presuppose the existence of an underlying ability that enables high performance on other tasks.

Therefore, a second area of assessments proposed by experts encompassed evaluations from computer science that target AI capabilities not included in human tests (bottom right in Figure 1.1). These are direct evaluations of systems on a task or a set of tasks provided in a standardised test dataset. The results of such assessments are typically held in publicly available leader boards. The tests sometimes refer to human performance on the task.

Third, experts considered real-world tasks involving a combination of capabilities for the assessment (top part of Figure 1.1). These tasks represent typical situations and scenarios occurring in education and work

and are, thus, instructive for AI's applicability in these settings. Such tasks can be found in some of the education tests discussed above. Although they target isolated capabilities such as reading or mathematics, these assessments often cover a mix of capabilities, including various aspects of language, reasoning and problem solving. Another source for complex, real-world tasks is certification and licensure occupational examinations. These tests include practical examples of typical tasks for a profession.

Taken together, this work showed that there are numerous capabilities and tests that can be used for assessing AI. A comprehensive assessment of AI must bring together different measurement approaches.

The second stage of the project

In its second stage, the AIFS project explored in further depth the three sources of assessments described above. It conducted exploratory assessments of AI capabilities with both education tests and complex occupational tasks. It commissioned experts to develop approaches to selecting and systematising direct evaluations used in AI research. The following sections summarise this work.

Exploring the use of education tests for collecting expert judgement on AI

The project continued the exploration of assessing AI on education tests with expert judgement set out in the pilot study in 2016. The aim was to test the feasibility of these assessments and further refine their methodology. The exploratory work addressed several broad methodological questions:

- What are the best methods for collecting expert judgement on AI with education tests (i.e. with regard to number of experts, method of expert knowledge elicitation, instructions for rating)?
- Does the approach produce robust measures with respect to different capabilities (i.e. capabilities that have been the focus of AI research, such as language processing, versus those that have received less research attention, such as quantitative reasoning at the time of the PIAAC numeracy assessment)?
- Can one reliably reproduce the assessment to track progress in AI capabilities over time?

The project addressed these questions with two exploratory studies. In 2021, it carried out a follow-up to repeat the pilot assessment of AI capabilities with PIAAC (OECD, 2023[4]). The purpose of this follow-up study was twofold. First, it aimed to track progress with respect to AI's literacy and numeracy capabilities since 2016. This was for both the substantive interest in the result and to inform the project about the feasibility and necessary frequency of future updates. Second, it attempted to improve the methodology of the assessment by applying more structured methods of expert knowledge elicitation.

The results of the follow-up study revealed some additional areas for improvement. In the numeracy assessment, experts' ratings of AI's capabilities strongly diverged. This had to do with the fact that the numeracy domain included a more diverse set of tasks (e.g. reading tables, processing images, interpreting graphs) and that experts had different assumptions of how a system should address this task diversity. While some evaluated the ability of a single system to perform all different tasks at once, others assumed narrow systems dealing with specific types of tasks. In other words, experts were uncertain about the generality of the hypothetical system being evaluated.

The results from the PIAAC numeracy assessment led the project to a careful consideration of how the rating task is presented to experts. In March 2022, experts were invited to reflect on a more clear-cut description of the rating instructions. The input from this meeting was used to develop a new framing of the rating exercise. In September 2022, four experts with expertise in quantitative reasoning of AI were invited to complete the numeracy assessment using the new framing. The goal was to test the new rating exercise and gather specialised expertise on the domain that may help better understand the challenges leading to disagreement.

In June 2022, the project extended the assessment to collecting experts' ratings on potential AI performance on PISA science questions. This new study aimed at testing a different approach for expert knowledge elicitation for the purposes of the project. Instead of working intensively with a small group of familiar experts, the study carried out a one-time online survey of a larger group of computer scientists. The goal was to gauge the feasibility of engaging more experts in terms of time, and human and financial resources, and to compare the robustness of these results to those relying on fewer experts.

Exploring the use of complex occupational tasks for collecting expert judgement on AI

The project has extended the rating of AI capabilities to complex occupational tasks taken from tests used to certify workers for different occupations. These tests present practical tasks that are typical in occupations, such as a nurse moving a paralysed patient, a product designer creating a design for a new container lid, or an administrative assistant reviewing and summarising a set of email messages. Such tasks are potentially useful as a way of providing insight into the application of AI techniques in the workplace.

The inherent complexity of these tasks makes them different from the questions in education tests used in previous assessments. Occupational tasks require various capabilities, take place in real-world unstructured environments and are often unfamiliar to computer scientists. Consequently, the project had to develop different methods for collecting expert ratings of AI with such tasks. This effort was guided by two main questions:

- What are the best methods for collecting expert ratings of AI and robotics performance on complex occupational tasks (i.e. instructions for rating, framing of the rating exercise)?
- Does the approach produce robust measures with respect to different occupational tasks (i.e. in terms of description of tasks, task complexity, types of capabilities required)?

In July 2022, a first exploratory study asked 12 experts to rate AI's ability to carry out 13 occupational assignments. A subsequent workshop discussed the results and the methodology of this assessment. The study aimed to collect first insights into the challenges that experts face in rating performance on the tasks and to develop corresponding solutions. The 13 occupational tasks were selected to cover diverse capabilities (e.g. reasoning, language and sensory-motor capabilities), occupations and working contexts. The materials describing the task varied in length and detail. This helped explore how different conditions for rating affect the robustness of the results.

In September 2022, a follow-up evaluation of the same tasks was conducted to test a new framing of the rating exercise. Experts were asked to rate potential AI performance with respect to several, pre-defined capabilities required for solving the task. The expectation was that linking occupational tasks to specific capability requirements would help experts abstract their evaluations from the concrete work context and focus more on general technological features needed for performing the task. A subsequent workshop with the experts elaborated the advantages and limitations of this approach.

Box 1.1. Types of AI measures discussed in the report

AI research utilises a broad range of tools for assessing performance, including benchmarks, tests, datasets, validations, performance metrics, evaluation frameworks and competitions, among others. Often, these terms are not used consistently across the research landscape, creating confusion amongst experts and non-experts alike. In this report, the term "measure" refers to any tool or method that evaluates AI performance.

These measures are categorised into *direct* and *indirect*. Measures that are constructed from results of standardised tests of AI performance are direct. In Chapter 6, Cohn and Hernández-Orallo refer to these measures as “evaluation instruments”, in line with their previous work on AI evaluation. By contrast, measures resulting from experts’ second-hand evaluation of the results of direct tests are indirect.

Direct measures

Direct measures are quantitative tools that assess specific performance characteristics of an AI system, generally under controlled or standardised conditions. They include:

- Benchmarks: These are standardised tests designed to measure the speed or quality of an algorithm’s performance. Example: ImageNet for visual recognition tasks (Deng et al., 2009[5]).
- Datasets: Collections of data used to train and test AI models. Example: MNIST dataset (Modified National Institute of Standards and Technology dataset) for handwritten digit recognition (Li Deng, 2012[6]).
- Competitions: Events where various AI models compete against each other in predefined tasks. Example: RoboCup for robotic soccer (RoboCup, 2023[7]).

Indirect measures

Indirect measures involve second-hand evaluations, often dependent on expert judgement or collation of existing research, aimed at gauging an AI system’s effectiveness or potential. They are ultimately based on direct measures. Indirect measures include:

- Expert Surveys: Questionnaires or interviews with experts who provide evaluations of an AI system’s capabilities. Example: The AI Index's annual report (Maslej et al., 2023[8]).
- Meta-Analyses: Comprehensive reviews of existing literature and datasets to provide an overarching view of AI performance. Example: Review of recent advances in natural language inference (Storks, Gao and Chai, 2019[9]).
- Validations: These are expert reviews or third-party assessments that evaluate the reliability and effectiveness of an AI system in real-world or simulated conditions. Example: Validation of AI in medical diagnostics by the Food and Drug Administration (FDA) (see Note).

Note: The FDA is providing a list of AI-enabled medical devices marketed in the United States under: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices#resources (accessed on 06 October 2023).

Exploring the use of direct AI measures

As a result of the challenges encountered in the use of expert judgement, the project began an initial exploration of the possible use of AI measures stemming from direct evaluations of AI systems (see Box 1.1). Hundreds of such evaluations exist, so-called benchmark tests, organised by research,

industry or other groups interested in promoting AI technology. These evaluations vary with respect to quality, complexity, purpose and the AI capabilities they target. They are also not systematised in a way that allows evaluations of higher-order capabilities or comparisons to human skills. The project thus needed to solve three methodological issues:

- How can one select good-quality measures among existing direct measures of AI?
- How should one categorise selected direct measures according to the AI capabilities they assess?
- How can one synthesise the results of direct measures into a few AI capability measures that allow for comparisons to human skills?

The project commissioned experts to work on each of these questions:

First, Anthony Cohn and José Hernández-Orallo developed a method for selecting existing measures for the assessment. This is a set of facets that describes and evaluates existing evaluation instruments for AI. On each facet, the researchers defined preferable characteristics of AI evaluation instruments. That is, AI evaluations with "desirable" values on many facets would be potentially useful for assessing the state-of-the-art of AI technology. The authors tested the rubric of facets on 36 benchmark tests from different AI domains.

Second, Guillaume Avrin, Swen Ribeiro and Elena Messina presented evaluation campaigns of AI and robotics at the French National Laboratory for Metrology and Testing (LNE) in France and the National Institute of Standards and Technology (NIST) in the United States. They proposed an approach for systematising these AI evaluations according to AI capabilities and identifying capabilities that have not been subject to evaluation.

Third, Yvette Graham reviewed major benchmark tests in the domain of Natural Language Processing (NLP). She then developed an integrated measure of natural language capabilities based on the reviewed tests. The measure provides links to expected human performance on the benchmark tests to enable AI-human comparisons across different language domains.

Outline of the structure of the report

This report is organised as follows:

Chapter 2 by *Abel Baret, Nóra Révai, Gene Rowe* and *Fergus Bolger* presents the evolution of methods the project used to collect expert judgement on AI capabilities from computer scientists and other experts. The chapter provides an overview of key methods of expert knowledge elicitation. The authors then describe the methodology used across the exploratory studies, including the different approaches to collect and analyse assessments from experts, the number of experts involved and the framing of tasks for experts. The chapter concludes with a discussion of the opportunities and challenges of using expert judgements and offers points of consideration for the project.

Chapter 3 by *Mila Staneva, Abel Baret et al.* presents the exploratory work on the use of education tests for collecting experts' assessments on AI. Three exploratory studies are described – the pilot study with PIAAC of 2016, its follow-up and the study using PISA. The chapter presents and compares the methodologies of these studies and discusses their results. It focuses on identifying best practices in collecting expert evaluations on AI with tests developed for humans.

Chapter 4 by *Mila Staneva, Britta Rüschoff* and *Phillip L. Ackerman* discusses the usefulness of complex occupational tasks for collecting expert judgement on AI and robotics capabilities. These tasks stem from occupation certification and licensure examinations and reflect typical situations and scenarios in the workplace. The chapter provides an overview of occupation examinations used in German vocational education and training and in the United States. It then describes in more depth 13 example tasks selected for an exploratory assessment of AI and robotics performance in occupations.

Chapter 5 by *Margarita Kalamova* presents two exploratory assessments of AI and robotics performance on complex occupational tasks. These studies test out and compare different methods for collecting expert judgement with complex tasks from occupational examinations. The chapter presents the results of these studies and discusses strengths and weaknesses of their approaches. It concludes by describing how assessments using occupational tasks will be used in overall project methodology.

Chapter 6 by *Anthony Cohn* and *José Hernández-Orallo* proposes a method for describing the characteristics of AI direct measures to guide the selection of existing measures for the assessment. Some of these characteristics have preferred values that identify good-quality direct measures to use for describing AI capabilities and their progress over time. The chapter describes the evaluation framework and tests it on a sample of 36 AI direct measures that cover different domains of AI.

Chapter 7 by *Guillaume Avrin, Elena Messina* and *Swen Ribeiro* provides an overview of the direct measures of AI resulting from the numerous evaluation campaigns organised by NIST and LNE. Evaluation campaigns in AI refer to comprehensive, structured and organised efforts to assess the performance of particular AI systems against objective quantitative criteria. The chapter systematises these campaigns according to the capabilities they address and identifies capability domains that have not yet been evaluated.

Chapter 8 by *Yvette Graham*, edited by *Nóra Révai*, reviews existing benchmark tests in the field of NLP and synthesises their results into a conceptual model for assessing AI language competence. The model provides a straightforward way for evaluating state-of-the-art AI performance in key NLP sub-domains. It also allows for comparing AI and human language competences.

Chapter 9 by *Stuart Elliott* summarises the results of the explorations described in this volume. It then outlines how these insights will be used for developing AI measures for key AI capabilities in the subsequent stage of the AIFS project. Concretely, the chapter explains how expert judgements on AI and existing measures from direct AI evaluations can offer a complementary approach for periodically measuring AI capabilities and comparing them to human skills.

References

Deng, J. et al. (2009), "ImageNet: A large-scale hierarchical image database", *2009 IEEE Conference on Computer Vision and Pattern Recognition*, https://doi.org/10.1109/cvpr.2009.5206848. [5]

Elliott, S. (2021), "Building an assessment of artificial intelligence capabilities", in *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, Paris, https://doi.org/10.1787/01421d08-en. [3]

Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/9789264284395-en. [2]

Li Deng (2012), "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]", *IEEE Signal Processing Magazine*, Vol. 29/6, pp. 141-142, https://doi.org/10.1109/msp.2012.2211477. [6]

Maslej, N. et al. (2023), *The AI Index 2023 Annual Report*, https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (accessed on 6 October 2023). [8]

OECD (2023), *Is Education Losing the Race with Technology?: AI's Progress in Maths and Reading*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/73105f99-en. [4]

OECD (2021), *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/5ee71f34-en. [1]

OECD (2021), *The Assessment Frameworks for Cycle 2 of the Programme for the International Assessment of Adult Competencies*, OECD Skills Studies, OECD Publishing, Paris, https://doi.org/10.1787/4bc2342d-en. [11]

OECD (2012), *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*, OECD Publishing, Paris, https://doi.org/10.1787/9789264128859-en. [10]

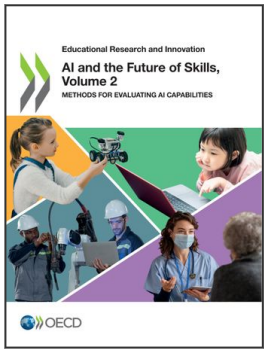
RoboCup (2023), *RoboCup Standard Platform League*, https://spl.robocup.org/ (accessed on 6 October 2023). [7]

Storks, S., Q. Gao and J. Chai (2019), "Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches". [9]

Notes

[1] In the following, the term “AI” will refer to both AI and robotics applications.

[2] The First Cycle of PIAAC (2011-17) assesses problem solving in technology-rich environments. It is defined as the ability to use “digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks” (OECD, 2012[10]). The focus is not on “computer literacy”, but rather on the cognitive skills required in the information age. The Second Cycle, which is under way, assesses adaptive problem solving instead. This is the ability of problem solvers to handle dynamic and changing situations, and to adapt their initial solution to new information or circumstances (OECD, 2021[11]).



From:

AI and the Future of Skills, Volume 2

Methods for Evaluating AI Capabilities

Access the complete publication at:
https://doi.org/10.1787/a9fe53cb-en

Please cite this chapter as:

Staneva, Mila (2023), "Overview", in OECD, *AI and the Future of Skills, Volume 2: Methods for Evaluating AI Capabilities*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/2a4e19b9-en