9 Project goals, constraints and next steps

Stuart Elliott, OECD

The AI and the Future of Skills (AIFS) project aims to provide indicators of AI capabilities for policy makers and other non-specialists. This requires a feasible approach that links AI to education and work and makes use of the substantial resources of available AI benchmarks and formal evaluations. The project's first methodology report explored skill taxonomies and tests to measure AI capabilities. This report focused on using human tests related to education and work for assessing AI, as well as on how to synthesise available tests from AI research. This chapter synthesises the implications of this exploratory work and outlines the next steps for the project. This subsequent work will consist in developing scales for different AI capabilities that integrate measures drawn from several sources and that link to the capabilities used in occupations.

The AI and the Future of Skills (AIFS) project aims to develop indicators of artificial intelligence (AI) capabilities that policy makers and other non-specialists can use to understand the implications of AI for work and education. The project's first methodology report explored a variety of skill taxonomies and tests that could be used for measuring AI capabilities. It concluded the eventual approach would need to bring together several different sources of information. These include measures drawn from human tests and those developed for AI, as well as those developed to test isolated capabilities and those developed to test more complex tasks. The project's subsequent development work has focused on exploring the use of human tests related to education and work for assessing AI, as well as ways to synthesise available tests from AI research. This work has been described in the preceding chapters of this volume. This chapter synthesises its implications and outlines the next steps for the project.

Potential sources of information about AI capabilities

As discussed in this report, there are two basic types of potential information about AI capabilities:

- *Expert judgement* about the current strengths and limitations of AI capabilities and about the existing instruments to measure those capabilities.
- *Direct tests* of AI systems where systems are applied to various kinds of tasks, producing success or failure. Examples of direct tests include:
  - *benchmarks* that provide a set of tasks for rating performance on a particular type of problem
  - *competitions* that enable multiple groups to develop systems to solve a particular (set of) test problem(s)
  - *formal evaluations* that provide a more intense analysis of the successes and failures of different approaches within a particular domain.

The original proposal for AIFS involved using expert judgement to evaluate AI capabilities with respect to specific tasks that could be compared to results for humans. The proposal grew out of a pilot study where questions from OECD's Survey of Adult Skills of the Programme for International Assessment of Adult Competencies (PIAAC) were used to gather expert judgements about corresponding AI capabilities (Elliott, 2017[1]). This approach collected expert judgements about the ability of current AI systems to answer individual PIAAC questions in the domains of literacy and numeracy. It then aggregated the ratings to compare the results for AI to the results of adult respondents.

This volume describes the project's recent work in exploring and refining the methodology for collecting expert judgements with human tests, as well as subsequent efforts to use the competing approach of direct AI measures. After reviewing explorations related to expert judgements on education tests and complex occupational tasks, this chapter describes explorations related to direct measures. The lessons from these three explorations to date are then used to describe an integrated conclusion about the potential use of these sources of information.

Exploration of techniques to elicit expert judgement

AIFS was intended to focus on use of expert judgements related to specific test questions. Consequently, the project held an early meeting to explore techniques to elicit expert knowledge related to judgements about quantitative values in complex domains. As described in Chapter 2, this meeting suggested several different techniques from the literature to select experts, obtain judgements from them, provide feedback and updates on those judgements, and aggregate the results, sometimes with various weighting approaches.

Chapter 3 describes the project's first expert judgements on specific test questions – a set of five-year updates of the feasibility of AI systems answering the OECD's PIAAC test questions in adult literacy and

numeracy. In literacy, the results were reassuring about the method, showing an increase in expected AI performance between the pilot study in 2016 and the follow-up study in late 2022 (consistent with AI progress related to natural language processing). They also showed a narrowing of the dispersion in responses across experts, and consistency in the ratings of experts who provided ratings in both years (OECD, 2023[2]).

In contrast, the numeracy results were not plausible. They showed a small decrease in expected AI performance compared to the 2016 results, a widening of the dispersion in responses across experts and inconsistency in the ratings of some experts who provided ratings in both years (OECD, 2023[2]). Qualitatively, the experts also expressed uncertainty about how to think about the meaning of their ratings with respect to the numeracy questions. Many said that any specific type of numeracy question could be answered if an AI system were developed for that type of question. It followed that the underlying problem was in understanding the domain and how many of the question types could be specified in advance to help develop an AI system.

The results on the PIAAC numeracy questions led the project to a careful consideration of the way the rating task had been presented to the experts. Since the experts were uncertain how to think about the level of generalisation needed by an AI system to answer the numeracy questions, the project team developed a new framing for the rating question (Chapter 3). This new framing involved an initial presentation of the test domain with descriptions and examples of the questions included on the test. It then asked experts to imagine an AI system based on current techniques that could be developed and trained for that domain. Finally, it asked the experts whether their imagined AI system could answer the remaining questions on the test.

The project team first tried out the new framing with science questions from the OECD Programme for International Student Assessment (PISA). In response, experts noted that the framing approach provided a better way of understanding what they were being asked to rate. The team also obtained ratings from more experts for both the PIAAC numeracy questions and the PISA science questions, using the new approach for framing the rating task.

For the PISA science questions, the team also tried to collect a substantially larger sample than the initial 10-12 experts. This would allow tighter statistical estimates. Thus, instead of working intensely with small groups of familiar experts, the team tried to collect judgements from many experts through an online survey. Chapter 3 describes the efforts along these lines and the conclusions related to the feasibility of obtaining larger samples.

The general findings from this initial work with the PIAAC and PISA test questions are as follows:

- Expected AI performance on these test questions is somewhere in the middle of the human performance distribution (as of late 2021 through mid-2022). Most experts believe many questions are easy for AI, but also rate many questions as not yet solvable.
- It is surprisingly difficult to obtain clear expert judgements that current AI techniques either can or cannot answer a particular question from these tests. Part of the difficulty relates to specifying the conditions that indicate what it means for an AI system to “be able to answer” a question from such a test. In addition, there is substantial disagreement across experts in their ratings for each test question, even though there is substantial agreement in their qualitative statements about AI capabilities.
- Expert ratings of AI performance on the PISA science questions closely match the performance of GPT-3.5 on these same questions (Chapter 3). GPT-3.5 was released by OpenAI in autumn 2022, with ChatGPT, a chatbot based on this model, released at the end of November. Experts completed the PISA assessment in summer and autumn 2022. Thus, despite the difficulties in obtaining quantitative agreement among experts, the aggregated judgements yield valid results regarding current state-of-the-art AI capabilities.

- There are practical limits to using this expert judgement process to obtain ratings about test questions. The pool of qualified experts who might be recruited to provide ratings is relatively small – probably several hundred worldwide. This is because the rating task is time-consuming, requiring several hours of work for the questions from each of the different tests.
- The quality of results from smaller expert groups using the behavioural approach was comparable to that obtained from larger groups using the mathematical approach. Meanwhile, the small-group assessments also provided important qualitative information and proved to be more feasible. This was due to the smaller number of people involved and the more intense interaction required.

Initial ratings of occupational performance tasks

The project team also extended the rating process (Chapter 5) to look at complex performance tasks (Chapter 4) taken from tests used to certify workers for different occupations. These tests present practical tasks typical for the occupation, such as a nurse moving a paralysed patient, a product designer creating a design for a new container lid, an administrative assistant reviewing and summarising a set of e-mail messages or a cosmetologist performing a manicure.

The initial evaluation asked experts to rate the feasibility of AI performing the entire task, as well as several individual subtasks. In general, that rating task appeared to be feasible. However, as with the adult numeracy test, experts were unsure how to think about rating the task. For the occupational tasks, the rating difficulty related to how much the task could be adapted and the underlying capability requirements for each performance task. As a further complication, AI experts were sometimes unfamiliar with the occupational tasks. This made it hard for them to anticipate typical difficulties in the work contexts where the tasks are performed.

In explaining their ratings, the experts often described various capabilities required by the task but not yet sufficiently advanced in AI systems. A follow-up evaluation of the same tasks asked experts to assess AI on several separate capabilities required for the task. This was intended to collect more nuanced information on AI performance on the tasks in a way that is apparently more familiar to AI experts.

This rating exercise was only a partial success for several reasons. First, values on the capability scales were not described in concrete terms. Second, there was confusion about the difference between the ratings describing AI capabilities versus the ratings describing the performance level of the capabilities required by the tasks. However, the experts generally agreed it was helpful to think in terms of different capabilities required for the task when evaluating AI's potential performance on the task.

This exploration highlighted the inherent complexity of work tasks, which involve numerous individual capabilities. This complexity makes it difficult to provide ratings of AI's capabilities in relation to the task. To do so requires judgements of all the required capabilities individually, as well as their combination. As a result, working with such tasks may be more useful for understanding the potential application of AI techniques for different types of work tasks than for gathering expert judgements about the current level of AI capabilities.

Explorations of direct measures of AI performance

The AIFS project initially proposed to use expert judgements on the ability of AI to answer questions from human tests. This proposed methodology anticipated that experts would likely use their knowledge of AI performance on existing direct measures to inform their judgements. However, it did not anticipate using the direct measures themselves to construct the project's indicators of AI capabilities.

During the initial exploratory phase, the project team substantially re-evaluated the potential role of direct measures on the project. This occurred for several reasons:

- AI experts repeatedly noted their concerns that human tests are designed to measure differences in capabilities that are important for decisions about people. These tests would not necessarily reflect the key differences in capabilities that matter to AI.
- When describing current AI capabilities, experts naturally described direct measures that are available and used by the field. They often noted specific limitations about those measures that are known in the field and that researchers are attempting to fix.
- The practical limits on using expert judgement heightened the importance of finding a more robust source of information about AI capabilities. With potentially thousands of measures available across the different subfields of AI, direct measures offer a substantial resource for the project.

As a result of this re-evaluation, the project began exploring the possible use of direct measures. This initial work involved the three efforts described in Chapters 6-8. These explorations suggest the following:

- There is a large number of direct measures and many follow rigorous protocols. However, they are highly scattered. There is no consistent taxonomy to categorise them, and they differ in nature and quality.
- The landscape of direct measures evolves rapidly. When AI systems can successfully perform a benchmark, it is no longer relevant. Meanwhile, new ones appear constantly to reflect state-of-the-art research and development (R&D).
- Human comparisons do not exist in most cases. When they do, they often compare AI performance to small samples, either random ones or specific ones such as human experts.
- Direct AI measures do not cover all human skills (understandably). It can be difficult to find a correspondence between these and human skills because the AI systems may focus on specific components or applications that are irrelevant for humans.

As a result, it is difficult to synthesise direct AI measures and create aggregate indicators of AI performance that are valid over time. In addition, it is difficult to compare human and AI capabilities based on direct measures.

The project is working with other researchers to explore ways of connecting detailed task descriptions to existing direct measures. In addition, they are developing new tasks to illustrate and potentially assess aspects of capabilities beyond current AI techniques.

One of the themes of this work so far is the challenge of synthesising results across numerous potential measures and then relating that synthesis to human performance.

Implications of recent work for the project's approach

After exploring these approaches, it is becoming increasingly clear that both expert judgements and direct measures of AI are necessary and, indeed, cannot be entirely separated.

On the one hand, results from current direct evaluations will not exist for performance levels clearly below or above the current state-of-the-art of AI. In contrast, expert judgement about performance on tasks that are too easy or hard for current systems should be easy to obtain and relatively consistent across experts.

On the other hand, expert judgements related to current areas of R&D are likely to be limited by experts' awareness of the most recent developments in AI systems. This is likely to lead to a lack of consensus across experts. In contrast, direct results will be available precisely for those areas that are the focus of current R&D, and will provide at least partial answers about AI performance in those areas.

This argument suggests that direct measures are indeed useful for understanding AI capabilities in areas of current research. However, even here, the direct measures will rarely stand on their own. Instead, expert judgement will be needed to choose among the many direct measures available, describe the limits in the types of performance that the measures reveal, and then develop a meaningful synthesis of those measures with respect to a broader capability.

This discussion implies that the relevant type of information for any given performance level for a capability will change over time. Initially, when the level of performance is too difficult for AI to attempt, there will be no direct measures. Expert judgement will then be the only source of information (i.e. there is no work with respect to that type of performance because it is too difficult). Later, when that type of performance becomes an active area of AI development, the primary source of information will become the direct measures that track that development process. However, these measures will need to be selected and integrated using expert judgement. The final step occurs when the problem of producing that level of performance is effectively solved and no longer an area for active research. At that stage, the field will no longer actively produce direct measures to demonstrate performance. Consequently, expert judgement will again provide the sole information about performance level.

Information needed about AI's implications for education and work

With a more realistic understanding of the constraints on gathering information about current AI capabilities, AIFS is considering the type of indicators relevant for highlighting differences or changes in AI capabilities that have implications for education and work. This section considers the types of education and work policy questions that indicators of AI capabilities should help answer. It then describes how AI indicators can address such questions by linking AI to human capabilities that are taught in education systems and used in the workplace.

Some major policy questions for indicators of AI capabilities

AI can potentially disrupt existing patterns of skill demand on the labour market and processes of skill development in education systems. Indicators of AI capabilities are crucial to help answer policy questions related to such potential education and work disruption:

- **Implications for curriculum**: How might new AI capabilities change the types of capabilities that people need to be prepared for work? What knowledge and skills should schools continue, stop and start developing? How will human and AI capabilities complement each other?
- **Implications for the goal of education**: What are the attitudes and values that remain or become important? How will new AI capabilities change the number and profile of people whose skills are below those of AI across essentially all capabilities used at work? What does that change imply for the role of education in preparing people for work and for adult life?
- **Implications for pedagogy**: How might new AI capabilities change the approach to teaching? How will teachers' work change?
- **Implications for the structure of education**: How might new AI capabilities shift the distribution of education across the lifespan, specifically with respect to the contrast between initial education and education later in adulthood, and with respect to formal and informal education?

For AI indicators to help answer such questions, AI and human capabilities will need to be compared in meaningful and accurate ways. They need to show how the roles of humans and AI will evolve as AI capabilities advance.

Linking AI and human measures of capabilities

One way to link indicators of AI capabilities to questions related to education and work is through the features used to describe occupations – their typical tasks and the skills, abilities and education they require. Such descriptions already exist and are widely used in planning and analysis for education and work. Systems that describe occupations for analysing the labour market – notably O*NET[1] in the United States and ESCO[2] in Europe – include measures related to skills or abilities, and activities or tasks, as well as educational qualifications. If indicators of AI capabilities can be naturally linked to some of these categories, it will be more straightforward to use the AI indicators to study AI's potential implications for work and education.

O*NET includes several complementary taxonomies related to work and workers. There is a high degree of overlap across separate taxonomies related to the categories of skills, abilities and activities. In each case, multiple scales relate to a few large clusters: language, reasoning and problem solving, sensory interpretation, motor control and social interaction. Probably any of these taxonomies would be a feasible basis for constructing a set of capability indicators for AI that could be linked to information available about education and work.

The AIFS project will make a pragmatic choice about working with one or more of these categories and then adjust as needed given feedback from computer scientists. At the current stage of development, the project will continue to refer to indicators of AI "capabilities". However, the scales ultimately used in these indicators could be more closely related to any one of these three different taxonomies that have been used to describe human work and workers.

Grouping AI capability indicators by their implications for education

When one compares topics covered by education to the capabilities needed at work, it becomes obvious that education focuses on developing only a portion of the necessary work capabilities. In the initial AIFS work with experts to analyse occupational performance tasks (Chapters 4-5), a portion of each task involves reasoning and problem solving. These could draw on professional instruction for the occupation (either from an academic or more vocational setting).

While the expert discussion did include these points, most of their analysis and conversation focused on some challenging capabilities not typically learnt in formal education (though they may be refined there). Such capabilities involve the situational awareness to understand the context of a workplace and identify tasks in that context that need to be performed; the common-sense knowledge and reasoning to understand how to connect and apply more abstract professional instruction to the complexity of a real workplace; and the sensory interpretation and physical movements necessary to perform actions involved with the task.

In considering the different types of capabilities needed in real work tasks, policy makers might distinguish among three types of human capabilities used at work that are addressed differently by education and training systems:

- *Basic capabilities*, like reading, writing and basic quantitative and scientific reasoning, are usually developed in formal education, often across the full population and in the younger grades.
- *Professional capabilities* include advanced reasoning in subjects like medicine, computer science or plumbing. They are also usually developed initially in formal education (either academic or vocational). However, they are typically developed only by subgroups in the population and usually by older students in later secondary or tertiary education.
- *Common capabilities* include understanding and using speech, reasoning about everyday situations, interpreting sensory information, moving one's body and manipulating objects, and

interacting socially. These capabilities are usually acquired developmentally and learnt without much formal instruction. However, they may be later refined with specialised professional training.

Because these different types of human capabilities are systematically related to education and training, there may be clear education policy implications if AI capabilities develop more quickly on one or two of these types of human skills.

- If AI progresses more quickly on common skills, there may be relatively more need for the basic and professional skills developed in formal education. Many people who primarily use common skills at work may need to further develop their basic and professional skills.
- Conversely, if AI progresses more quickly on basic and professional skills, there may be less need for the skills developed in formal education. The duration and approach of formal education may need to be substantially changed.
- Similarly, if AI progresses more quickly on basic skills than on expert skills, or vice versa, then the mix between these skills in formal education may need to be substantially changed.
- Substantial numbers of people are likely to be displaced from work only if AI progresses quickly on all three of these types of human skills.

Obviously, there are substantial distinctions within each of these three broad categories. The first two bring together a number of different capabilities that are important to distinguish with respect to education and training. It is not yet clear how to align AI capabilities to contrasts between skills developed within and outside education systems. The project will explore this possibility as the capability dimensions are defined.

Next steps for the project

The project team will integrate the insights gathered from the three exploratory efforts related to the use of education tests, occupational tasks and direct measures. It will then develop the assessment of AI capabilities and their implications for education and work. This section summarises the two major strands of the work to come: a systematic development of the indicators of AI capabilities and further exploration of approaches to analyse the implications of new AI capabilities for education and work.

Systematic development of the indicators of AI capabilities

The project is shifting to the systematic development of the indicators of AI capabilities. This development work is occurring within four broad domains related to language; reasoning and problem solving; sensory perception and motor control; and social interaction. Initially, indicators are being developed within each broad domain that make sense for describing current and future AI capabilities. These will then be linked to the occupational taxonomies included in O*NET and ESCO.

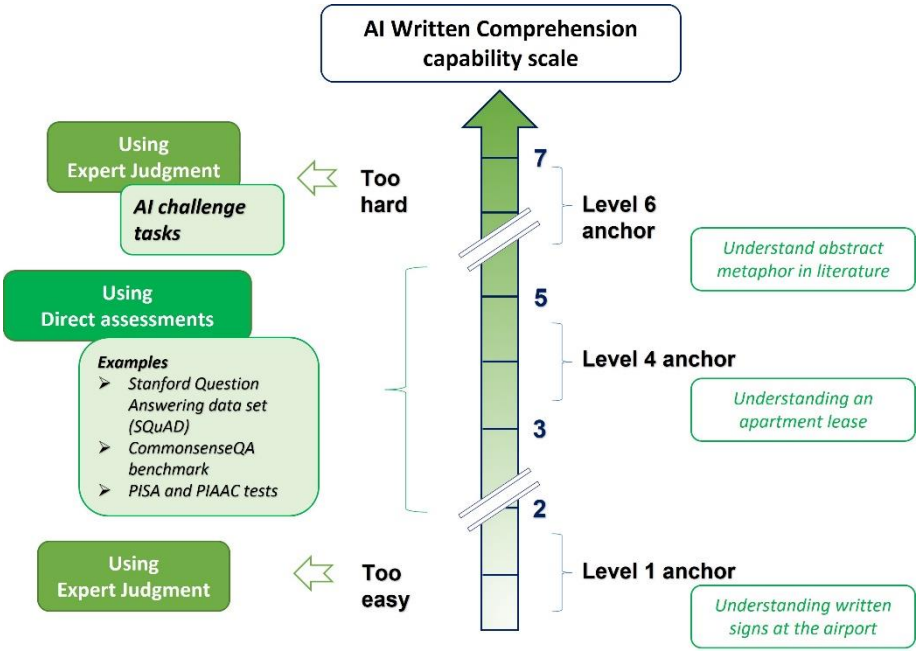
The selection of indicators to develop will reflect feedback from AI experts and policy makers about which taxonomy works best for translating AI capabilities to non-experts. It will also reflect their feedback on which scales from the occupational databases within that taxonomy are relevant to include for AI. In addition to the AI expert network helping to develop the indicators, the project will seek feedback from policy makers. Specifically, it will solicit opinions about which taxonomy and which specific scales will be most helpful in translating information about AI to its practical implications for education and work.

The project will explore the possibility of providing a high-level description of AI capabilities that maps onto the distinctions between basic, professional and common skills. To that end, it will consider whether the chosen capabilities can be grouped into those three categories. It will also look at whether it would be meaningful to create some aggregate indicators to communicate AI progress for each of these groups.

For each capability, the project is developing scales that reflect different levels of performance that are meaningful for AI. Figure 9.1 provides an example of a scale assessing the capability "written comprehension" to illustrate the concept. These scales will be described in terms of a set of anchoring tasks that are meaningful to both AI experts and non-experts. They will focus on anchors that receive consistent difficulty rankings from the AI experts. The scale development will include and illustrate higher levels of performance for each capability that are clearly beyond current AI techniques. This will provide concrete examples for non-experts of the current limits of AI capabilities and important milestones for future development. The higher performance levels will extend at least until the levels of performance that are consistent with expert human performance levels. They might even extend beyond those levels for capabilities where clear examples are available. Development of the anchoring tasks for the capability scales will involve a large, multidisciplinary group of AI experts.

For each of the resulting capability scales, the project will identify performance ranges that are clearly too easy or too difficult to be reflected in current direct measures. Within the range where direct measures provide information, it will develop some way of indicating the types of partial and full performance that current AI techniques provide, based on the information in available direct measures. Small teams with expertise related to each specific scale will develop these mappings from direct measures to the scale. Other experts who share the same expertise will peer review these mappings.

Figure 9.1. Conceptual scale reflecting AI performance levels



In addition to the direct measures used in the field, the project will consider the use of two other sources of example tasks. These may be important to illustrate some portions of the capability scales that are not well populated by existing direct measures:

- **Questions from human tests** may be useful in domains with well-developed tests and where the results may help illustrate important aspects of AI capabilities. In particular, education tests that play important roles in policy discussions about basic and professional skills – such as the PISA and PIAAC tests the project has already explored – are likely to provide useful perspectives on some AI capabilities.

- **Tasks that illustrate current AI challenges**, reflecting aspects of AI capabilities that are beyond current capabilities, are likely to be useful as a way of monitoring development in the field before being crystallised into benchmark tests.

These two additional sources of example tasks could be rated in two ways. They could use expert judgement as the project has done so far with the questions from PISA and PIAAC. Or they could obtain direct measures on these tasks through a competition or commissioned development of new AI systems.

The final aspect of developing indicators of AI capabilities will be translating the underlying scales to the corresponding human capabilities. This would aim to communicate how AI and human performance compares for each of the measured capabilities.

Further exploration of the implications of AI capabilities for work and education

Understanding the implications of AI capabilities for work and education will revolve around understanding how AI performance on the different capabilities can support humans across the full range of contexts, including education, work and daily life. The next steps of exploration will focus on finding ways to systematically understand the plausible implications of different AI capabilities on work and education. A later stage of the project will look at implications for AI in daily life beyond work and education.

With respect to work, the project is creating a sample of 25-100 tasks to represent different contexts, required skills and abilities, and component activities in jobs across the entire economy. The sample of work tasks will reflect the full diversity of the economy. At the same time, it will provide a set of concrete examples that can each be analysed in detail. (The wide range in the size of the sample reflects the current uncertainty in the size necessary to appropriately reflect the diversity of work tasks in the economy.)

The project plans for a group of AI experts and job analysts to study different sampled work task. This will determine which activities an AI system could perform. It will then propose ways to redesign the current task so a human can complete the AI-performed tasks with support of an AI system. This would make it possible to describe a transformed role for humans in each work task to illustrate the kind of transformation feasible with AI. The analysis of individual tasks would consider current AI performance levels for different capabilities, as well as several performance scenarios that AI could plausibly achieve in the next 5-20 years.

With respect to education, analysis of the potential use of AI capabilities will consider the types of human capabilities developed in formal education. It will also examine the possible use of AI systems that provide limited levels of those capabilities as support to humans. This would be akin to how calculators and computers have been incorporated into mathematical reasoning and the mathematics curriculum over the past several decades. The intent will be to anticipate how capabilities developed in formal education may be supported and transformed by AI systems that have partial or complementary versions of those capabilities. This would be important for AI systems with substantial levels of language and reasoning capabilities that might help develop student reasoning (as well as later professional reasoning at work) across a broad range of content areas.

The project will explore several paths to start addressing educational implications. First, it will work with a group of education researchers to explore how learning outcomes and educational standards might change in a scenario where AI can perform at a high level in a specific domain (e.g. science education). Second, it will examine how AI will affect the teaching profession as a specific occupation. Third, it will work with a group of experts (e.g. policy makers and curriculum developers) to explore non-traditional educational goals. These will go beyond preparing students for the labour market by, for example, taking a historical perspective or viewing through the lens of cultural minority communities.

The project will work with a group of education researchers and computer scientists to analyse a representative set of capabilities developed in formal education. They will describe the ways that humans

could work with the support of an AI system to perform these capabilities. The group will consider how this would transform the nature of the capability for humans and implications for the goals, curriculum and pedagogy in different subjects in formal education. The educational analysis for occupational tasks will consider current AI performance levels for the different capabilities, as well as several performance scenarios that AI could plausibly achieve in the next 5-20 years.

The AIFS project team will carry out exploratory work to develop feasible approaches for understanding implications for work, education and daily life. It will aim to identify a systematic approach across a sample of work tasks and educational topics. This work will be described in a later volume of this series of methodology reports.

References

Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/9789264284395-en. [1]

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. [2]

Notes

[1] Occupational Information Network (O*NET) by the U.S. Department of Labor, available at https://www.onetonline.org/ (accessed on 30 October 2023).

[2] Classification of European Skills, Competences, Qualifications and Occupations, available at https://esco.ec.europa.eu/en (accessed on 30 October 2023).