20. Building an assessment of artificial intelligence capabilities

Stuart W. Elliott, OECD

This chapter synthesises the expert contributions of the report and offers perspectives for building a comprehensive assessment of artificial intelligence (AI) capabilities. It compares and contrasts the contributions of psychologists and computer scientists along two dimensions: whether they focus on human or AI taxonomies and tests, and whether they test isolated capabilities or more complex tasks. The chapter argues that a more complete assessment of AI must bring together the different approaches. It illustrates this argument with an example in the area of language. Finally, the chapter offers next steps towards a systematic assessment of AI capabilities, which will allow for drawing fine-grained implications for work and education.

Introduction

Through the *Artificial Intelligence and the Future of Skills* (AIFS) project, the OECD is developing an approach to assessing the capabilities of AI and comparing them with human capabilities. The goal is to develop a comprehensive programme to measure these capabilities in a valid, reliable and meaningful way that policy makers can use to understand the implications of AI for education and work. AIFS is a six-year project, which will include an initial systematic assessment of AI capabilities and analysis of their implications. It will conclude with a proposed approach for an ongoing programme to assess AI capabilities at regular intervals.

This volume addresses questions related to identifying the AI capabilities that the project should assess, as well as tests that could be used to assess them. Based on a workshop in October 2020, the volume surveys a broad range of work in psychology and computer science that can provide relevant taxonomies of capabilities and assessments of them.

The papers in the volume make clear there are many resources the project can use to define a set of capabilities and associated assessments for measuring the capabilities of AI. Before the workshop, the AIFS project team hoped to identify a single comprehensive taxonomy that could be linked to an appropriate set of assessments to use for the project. Given the number of available taxonomies in psychology, this hope seemed potentially realistic. However, the experts who participated in the workshop and wrote papers in this volume put forward other thoughtful arguments. They suggest that AIFS could benefit from combining the complementary strengths of several different approaches rather than using only one.

The meeting discussion and this resulting report highlight two dimensions of difference that should be reflected in an assessment of AI capabilities: the contrast between human and AI taxonomies and tests, and the contrast between testing capabilities and tasks.

Human vs. artificial intelligence taxonomies and tests

As amply illustrated in Part I of this volume, rich resources in psychology reflect long research traditions to develop the conceptual and practical tools for cognitive assessment in humans, as well as animals. Furthermore, the efforts to assess AI capabilities in computer science have often started from these materials, as noted by Hernández-Orallo in Chapter 11, because of their broad coverage and availability. Indeed, the computer science community acknowledges the intellectual foundation and extensive materials provided by psychology. However, computer scientists also clearly state that human tests can be misleading and incomplete when used to assess AI. Many papers stress this point in Part III.

Computer scientists note that human tests focus on aspects of capabilities that are meaningful for assessing humans. However, these tests are not necessarily meaningful for assessing for AI. Assessments are always incomplete, focusing on only a sample of the capabilities of interest. They then assume the sampled capabilities also provide information about the critical unsampled capabilities needed for competent performance.

Because the cognitive capacities of humans and AI are different, assumptions about unsampled capabilities are different. Therefore, the sampled capabilities included on a test need to be different. As a result, AI needs to be assessed for things rarely considered for direct assessment in humans, such as common sense reasoning. This could lead to somewhat different taxonomies of the capabilities needed to consider for AI. Ultimately, this could lead to entirely new tests to assess those capabilities.

Testing capabilities vs. tasks

In Chapter 15, Avrin discusses the importance of assessing both isolated capabilities ("functionality benchmarks") and the performance of complete tasks ("task benchmarks") in evaluations of AI. Usually, the tasks are the priority. The task benchmarks are then chosen to reflect something wanted from an AI system in the real world.

However, tasks almost always require complex combinations of capabilities. An AI system may fail because of the inadequacy of either one of the capabilities or their integration. Assessing the individual capabilities provides a way to determine whether each one is sufficient for the task; the assessment of the task itself determines whether the separate capabilities function effectively together.

This contrast that Avrin describes on the AI side is richly illustrated across the different types of assessments on the human side.

On the one hand, assessments in psychology often attempt to isolate individual capabilities for assessment and avoid using tasks that will confound the contributions of several distinct capabilities. The process of separating the contributions of different individual capabilities with specially designed tasks is at the heart of the factor analytic tradition in psychology discussed by Kyllonen in Chapter 3.

On the other hand, many human assessments focus on authentic tasks of interest in human contexts and intentionally mix the full set of capabilities needed for those tasks. The occupational tests discussed by Rüschoff in Chapter 9 provide the clearest example of authentic tasks that require many separate capabilities to carry out. These tasks often involve not only cognitive capabilities related to language, reasoning and problem solving but also additional capabilities related to social interaction, sensory perception and psychomotor control. The educational tests discussed by Greiff and Dörendahl in Chapter 7 often aim at a middle range of complexity. They mix capabilities related to language, reasoning and problem solving but omit capabilities related to social interaction, sensory perception and psychomotor control that can be important in many work contexts.

Working with both dimensions

The four revolutions of Forbus

In Chapter 2, Forbus illustrates the two dimensions of difference – human vs. AI taxonomies and tests and testing capabilities vs. tasks – through four revolutions in AI.

The first three revolutions relate to the key categories of the human cognitive taxonomies: one can link learning, knowledge and reasoning directly to Carroll's 3-stratum model of human cognitive abilities, discussed by Kyllonen in Chapter 3, as general memory and learning, crystallised intelligence and fluid intelligence, respectively.

The fourth revolution – agency – relates to the complex way that humans can integrate capabilities. It is reflected, for example, in the complex tasks carried out in human jobs, as well as the basic developmental stages in children.

Forbus uses this revolutions framework to identify both recent successes and key missing aspects of current AI capabilities. Crucially, some missing aspects are ones that may not be typically assessed on the human side. These include the ability to learn from a single example, knowledge of one's personal experience and common sense reasoning. Agency then provides the example of the combination of capabilities that is still missing to carry out real-world tasks in context.

While illustrating the two key dimensions of difference, Forbus also highlights the larger motivation for the AIFS project: there are revolutions occurring or approaching in each of these four key areas of AI cognition

that will result in qualitative shifts in AI capabilities. The prospect of major improvements in AI capabilities underlines the importance of providing measures for the policy community that identify what capabilities are missing. These measures can provide an early warning system, identifying when those capabilities appear and offering guidance to their implications.

Combining the two dimensions

Figure 20.1 suggests a way to fit the two dimensions together in a framework for assessing AI capabilities. This figure provides an initial framework for synthesising the different taxonomies and tests discussed in this report.

The first dimension – differentiating between human and AI sources for assessment – is illustrated horizontally at the bottom of the figure. AI assessment approaches derived from human capability frameworks appear on the left, while assessment approaches focused on missing AI capabilities are on the right.

The second dimension – differentiating between testing separate capabilities and complex tasks – is illustrated vertically. Assessment of separate capabilities is at the bottom, while assessment of real-world tasks requiring use of multiple capabilities is at the top.

The boxes for human capability frameworks and real-world tasks reference some of the taxonomies that describe and categorise relevant capabilities and tasks, respectively, and that link to a variety of assessments. The box for missing AI capabilities differs from the other two boxes in listing capabilities that are "special cases" rather than listing frameworks. These special case capabilities are often missing in two senses: they are missing from AI's current capabilities and from many (but not all) of the capability frameworks and assessments used to describe humans. As a result, the missing AI capabilities are both important to assess and require extra effort to identify assessments focused on AI's unique challenges.

Figure 20.1. Sources for AI assessments

Filling in the details

The chapters provide a wealth of detail about the kinds of capabilities and tests that might go into each of the boxes in Figure 20.1.

Starting with the Human Capability Frameworks box, Chapters 3, 5 and 6 present taxonomies and tests for a set of isolated human abilities. Kyllonen in Chapter 3 outlines the comprehensive taxonomies developed to describe the full range of cognitive abilities, building on a rich assortment of associated tasks. These taxonomies have been extended by some researchers, and Kyllonen briefly discusses some of the work related to social-emotional, perceptual, psychomotor and other skills. Kyllonen's initial overview is then supplemented by more detailed discussions in some of the other chapters in Part II. De Fruyt in Chapter 5 discusses social and emotional capabilities, along with some tests developed to assess them. Woolley in Chapter 6 provides a detailed discussion of the components of social capabilities that allow groups to function effectively, along with some novel assessments of those capabilities.

These human taxonomies are well developed and provide an extensive set of human tests for the different isolated abilities that could potentially be applied to assess AI. Some areas appear to be less interesting for AI assessment because AI systems have already mastered the abilities or could be developed to do well on the test without the underlying capabilities of interest. It would be necessary to choose tests carefully, in some cases using existing tests as an inspiration to develop versions that would be more likely to produce valid results for AI.

Two chapters reside in the Human Capability Frameworks box but represent an attempt to identify frameworks and tests on the human side that might be developed to address some of the Missing AI Capabilities. Chokron in Chapter 4 describes the many domains and assessments used in neuropsychological evaluation in children. Research on the assessment of deficits of normal cognitive functioning in children raises the possibility of identifying assessments of some missing AI capabilities that are usually also missing from tests for adult humans. Similarly, Cheke and colleagues in Chapter 17 use approaches for testing basic capabilities in young children and animals to develop some tests for AI systems of these capabilities. This chapter is placed in Part III of the report because it has already made the move into a set of applications for assessing some missing AI capabilities, but it rests on a research foundation from human and animal psychology.

Moving completely over to the Missing AI Capabilities box, the various papers from computer scientists outline a set of examples of assessments that have been or could be carried out related to AI systems:

- Hernández-Orallo in Chapter 11 provides an overview of the different approaches.
- Avrin in Chapter 15 discusses a number of AI and robotics systems that have been formally evaluated at the Laboratoire national de métrologie et d'essais in France, including a number of individual capabilities. The chapter also discusses the assessment of complex tasks, which belong in the Real-World Tasks box.
- Graham in Chapter 16 describes the assessment of different components of natural language capability. She makes the case that the field of natural language processing has developed assessments that go beyond typical human assessments. They now focus on the specific challenges and current level of capability in available AI systems for natural language processing.
- In Chapter 14, Cohn describes a few of the competitions and benchmarks used to compare performance of AI systems in different areas. He notes how competitions and benchmarks often evolve to focus on performance levels that are almost but not yet attainable by the field.
- The papers by Davis in Chapter 12 and Granger in Chapter 13 illustrate how AI assessment can go awry. They provide surprising examples of "brittle" performance of AI systems where seemingly small task differences produce large differences in the results. The authors present these examples as cautionary tales. Yet the examples simultaneously illustrate assessment techniques that can identify such brittle performance, at least in some cases.
- Finally, Forbus in Chapter 2 outlines several possible strategies for assessing AI progress related to the different revolutions he describes.

These many efforts do not suggest an integrated framework for assessing AI with respect to the aspects of capabilities that are not well reflected on human tests. However, they indicate several different approaches that can be explored for doing so.

Moving up to the real-world tasks involving combined capabilities in Figure 20.1, several chapters consider educational or occupational tests. Many educational and occupational tests focus on isolated capabilities that would appear in the Human Capability Frameworks box of the figure. These include, for example, capabilities in skills related to reading or mathematics. However, the chapters on educational and occupational tests in this report largely focus on tests that require combinations of capabilities. These are tests inspired by complex tasks in the real world, occurring in the context of education or work.

In Chapter 7, Greiff and Dörendahl provide an overview of different educational tests, including both core domain and transversal skills. Each of the assessments focuses on a particular capability, like reading literacy or problem solving. However, all the assessment tasks discussed require a mix of capabilities, including various aspects of language, reasoning and problem solving.

Finally, three chapters provide an overview of occupational tests, and the complexity of the tasks that can sometimes be included in them. Ackerman in Chapter 8 argues for the benefits of assessing AI using domain-specific tests for different occupations that include assessment of both declarative knowledge and hands-on procedural knowledge. Rüschoff in Chapter 9 then introduces the testing programme in the German vocational education and training system. A detailed discussion of the framework includes examples of specific tests. Dorsey and Oppler in Chapter 10 provide an overview of the framework for understanding occupational tasks and worker capabilities included in the US Department of Labor's Occupational Informational Network, along with several examples of occupational tests.

Fitting the details into the framework

The framework in Figure 20.1 suggests the importance of bringing together different types of approaches to provide a more complete assessment of AI. Building on the examples discussed in the chapters, Figure 20.2 shows a partial example of what this might look like in the area of language. At the bottom of the figure are examples of tests for isolated language capabilities. One is a test of vocabulary noted by Kyllonen in Chapter 3 as an example from the human side. The other is the Winograd Schema noted by Cohn in Chapter 17, which was developed to assess AI's ability to identify difficult pronoun referents. Greiff and Dörendahl in Chapter 7 provide an example of a reading literacy task from the Programme for International Student Assessment. This involves understanding a reading passage, reasoning about it and providing a written answer. Finally, Rüschoff discusses a multi-hour work sample assessment for Advanced Manufacturing Technicians. This includes several oral discussions about a complex work task involving the construction of a functional module according to a set of technical drawings.

As illustrated in Figure 20.1, the chapters have gone beyond suggesting an assessment approach that combines several different types of assessments. They have also provided concrete examples of the types of assessment tasks needed for such an assessment.

In his reflections on the project in Chapter 18, Graesser raises the important point about how to integrate the various taxonomies suggested in the different chapters. In Graesser's terms, the provisional framework for a synthesis in Figure 20.1 is perhaps a "comprehensive" synthesis, where complementary aspects of different approaches are added together. However, the two key dimensions of difference suggested by the chapters also provide the initial ingredients for a synthesis that is more theoretically motivated. Over time, the OECD hopes this can move to a rough consensus related to the capabilities and types of assessment tasks that the AIFS project should include.

Figure 20.2. Assessing language with multiple tests

Next steps

This volume is only a starting point for the project. There is substantial work to do to build a specific set of assessments to provide policy makers with an understanding of AI capabilities and their implications for education and work.

Several chapters in this report described approaches for moving directly to assess implications of AI without going through the intermediate step of evaluating AI's capabilities. As described in the introduction, the research literature includes a number of efforts to take this more direct approach. While acknowledging the value of this work, the AIFS project is building a more substantial foundation related to understanding AI capabilities before moving on to their implications.

The project is motivated by the importance of developing a more robust and meaningful understanding of AI that can support policy makers in understanding its implications. This is particularly important with respect to the educational implications of AI – the primary motivator of the project – which require a fine-grained understanding of the way that human and AI capabilities will complement each other.

The next stage of the AIFS project will involve piloting the types of assessments described in this volume to identify how well they provide a basis for understanding current AI capabilities. This work will begin with intense feedback from small groups of computer and cognitive scientists who attempt to describe current AI capabilities with respect to the different types of assessment tasks. It will build the project's understanding of the types of assessment approaches that give a valid and reliable picture of AI capabilities. With more understanding of types of assessments to use, the project will expand the range of input to include a broader sample of experts who know about AI so that we can fully represent the field.

Baker and O'Neil in Chapter 19 and Graesser in Chapter 18 anticipate a number of challenges that will come in this next phase. Baker and O'Neil review a set of practical issues that must guide the process of gathering ratings from experts who know about AI, as well as the larger context and framing that experts must consider when providing their ratings. Graesser raises a key question about how one should evaluate a comparison of AI and human performance. He asks whether human performance should be defined as the standard for evaluating AI – as is often done in AI evaluations. Should an objective, ideal model be used instead? The project will need to address these questions in the next stages of development work.

The initial work has also given the project an appreciation for the range of empirical measures of AI capabilities – such as those discussed by Cohn in Chapter 14 – that could potentially provide some of the

assessments the project seeks to create. At this point, the usability of these empirical AI measures is an open question. Many experts are concerned that they are often too narrow to provide informative comparisons between humans and AI for people outside the field of computer science. As a result of these discussions, the project now expects the next steps to consider the potential value of these measures to the project.

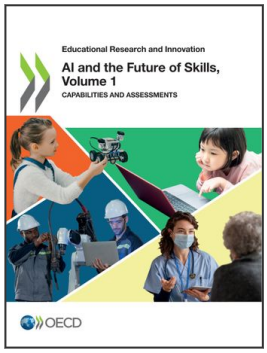
From assessment to implications

The project will ultimately translate assessments of AI capabilities to their implications for education and work. One part of this will involve a simple comparison of AI and human capabilities in different areas. This will aim to understand what aspects of different capabilities are still well beyond AI and how many people have those capabilities. In addition, the project will envision different scenarios for applying AI capabilities to the tasks in different occupations. This will be a way to understand how humans will begin to work with AI systems that have new capabilities and how human occupations will evolve, along with the educational preparation they require.

The last step of the translation process will be to develop ways to communicate the results of these assessments and analyses to policy makers and the general public. This will likely involve creation of a set of indicators across different capabilities and different work activities to communicate the substantive implications of AI capabilities in meaningful terms.

The development work is projected to continue through the end of 2024. At that time, a first systematic assessment of AI capabilities should be completed. A translation of that assessment to its implications for education and work, with a set of meaningful indicators that describe those results, is also expected to be finished. Of course, AI is advancing rapidly and a single assessment would quickly become outdated. The final stage of the development work will be to define a programme for regular updates to the assessment.

In the larger vision for this work, an ongoing programme of assessment for AI will add a crucial component to the OECD's set of international comparative measures that help policy makers understand human skills. The Programme for International Student Assessment (PISA) provides the link from education to skill, while the Programme for the International Assessment of Adult Competencies (PIAAC) provides the link from skill to work and other key adult roles. The AIFS project is now building a component that will relate human skills to the pivotal technology of AI, thereby providing a bridge from AI to its implications for education and work, and the resulting social transformations in the decades to come as AI continues to develop.



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:

Elliott, Stuart W. (2021), "Building an assessment of artificial intelligence capabilities", in OECD, *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/01421d08-en