



Educational Research and Innovation

AI and the Future of Skills, Volume 2

METHODS FOR EVALUATING AI CAPABILITIES



Educational Research and Innovation

AI and the Future of Skills, Volume 2

METHODS FOR EVALUATING AI CAPABILITIES

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

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Note by the Republic of Türkiye

The information in this document with reference to “Cyprus” relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Türkiye recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Türkiye shall preserve its position concerning the “Cyprus issue”.

Note by all the European Union Member States of the OECD and the European Union

The Republic of Cyprus is recognised by all members of the United Nations with the exception of Türkiye. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

Please cite this publication as:

OECD (2023), *AI and the Future of Skills, Volume 2: Methods for Evaluating AI Capabilities*, Educational Research and Innovation, OECD Publishing, Paris, <https://doi.org/10.1787/a9fe53cb-en>.

ISBN 978-92-64-88932-3 (print)
ISBN 978-92-64-82429-4 (pdf)
ISBN 978-92-64-42035-9 (HTML)
ISBN 978-92-64-81732-6 (epub)

Educational Research and Innovation
ISSN 2076-9660 (print)
ISSN 2076-9679 (online)

Photo credits: credits: © Shutterstock/LightField Studios; © Shutterstock/metamorworks; © Shutterstock/Rido; © Shutterstock/KlingSup.

Corrigenda to OECD publications may be found on line at: www.oecd.org/about/publishing/corrigenda.htm.

© OECD 2023

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

Foreword

Artificial intelligence (AI) has emerged as a significant area of development. Its integration into various sectors necessitates a comprehensive understanding of its capabilities, especially in relation to human skills. The AI and the Future of Skills (AIFS) project by the OECD's Centre for Education Research and Innovation (CERI) has undertaken this task, aiming to provide a methodological framework for assessing and comparing AI capabilities to human skills. This framework should provide a basis for informed discussions on AI's impact on education, work and society.

The project has undergone two phases of developing a rigorous approach to assessing AI's capabilities. The first phase focused on identifying relevant AI capabilities and the tests best suited to evaluate them. Leveraging insights from various fields including computer science, psychology and education, the project offered a multi-disciplinary perspective on the challenges and prospects of assessing AI.

The second phase, the focus of this report, further refines the methodology of the assessment. It encompasses a range of exploratory AI evaluations to identify most promising practices for systematically and periodically assessing AI. These explorations are threefold. First, by assessing AI capabilities with OECD's education tests using expert judgement, the project explored ways to understanding AI's progress in competencies that are traditionally human – competencies in reading, mathematics and science. Second, the project asked experts to rate AI on real-world occupational tasks, such as those encountered in nursing or product design, to provide critical insights into AI's application potential. By situating AI within these occupational contexts, we gain a clearer picture of its impending impact on the economy. Third, the project considered the vast and evolving benchmarks available in AI research that result from direct assessments of AI systems.

These methods, while promising, are not without their challenges. This report underscores the difficulties in solely relying on expert judgements to evaluate AI. While expert input is valuable, achieving consensus, particularly in novel domains, can be challenging. Moreover, the variability in AI applications and the intricacies of real-world tasks suggest the need for diverse evaluation metrics. Therefore, the project decided to integrate both expert judgements and direct AI measures in its subsequent phase to provide a thorough and balanced evaluation. This integrative approach aims to provide decision-makers with a nuanced understanding of AI's capabilities.

The next project phase intends to produce an integrated assessment framework for AI. This will contain a set of key AI indicators that can serve as reference points for various stakeholders. These indicators, informed by a combination of expert input and direct assessments, will offer guidance for policy formulation and implementation.

As AI continues to evolve, having a clear framework to understand its capabilities becomes crucial. The AIFS project's efforts contribute to this understanding, laying the groundwork for informed decisions in education and employment sectors. This work reflects OECD's commitment to producing rigorous, evidence-based insights that can inform decision-making in the context of AI's continued growth and integration into various sectors.

Acknowledgements

This publication was planned and developed by the OECD's Artificial Intelligence and Future of Skills project team – Stuart Elliott (Project lead), Mila Staneva, Margarita Kalamova, Abel Baret, Nóra Révai, Sam Mitchell, Marc Fuster-Rabella and Aurelija Masiulyté. The report was prepared for publication by Mila Staneva and Aurelija Masiulyté.

This publication would not have been possible without the invaluable contributions of the renowned computer scientists and psychologists who are supporting the project.

Firstly, we would like to express our gratitude to the experts who participated in the assessments or provided advice (in alphabetical order): Phillip L. Ackerman, Guillaume Avrin, Chandra Bhagavatula, Joseph Blass, Fergus Bolger, Jill Burstein, Salvador Carrión Ponz, Anthony G. Cohn, Vincent Conitzer, Ulises Cortes, Pradeep Dasigi, Ernest Davis, Angel de Paula, Marie desJardins, Kenneth D. Forbus, Carlos Galindo, Janice Gobert, Jordi González, Arthur C. Graesser, Yvette Graham, Fredrik Heintz, Jim Hendler, Daniel Hendrycks, José Hernández-Orallo, Jerry R. Hobbs, Lawrence Hunter, Juan Izquierdo-Domenech, Maria Juarez, Aina Juraco Frias, Ryota Kanai, Aviv Keren, Rik Koncel-Kedziorski, Patrick Kyllonen, David Leake, Bao Sheng (Aiden) Loe, Fernando Martinez-Plumed, Aqueasha Martin-Hammond, Cynthia Matuszek, Elena Messina, Antoni Mestre Gascón, Ángel Aso-Mollar, Jose Andres Moreno, Constantine Nakos, Taylor Olson, Rebecca J. Passonneau, Swen Ribeiro, Carolyn Rose, Gene Rowe, Vasile Rus, Britta Rüschoff, Vijay Saraswat, Areg Mikael Sarvazyan, Brian Scassellati, Wout Schellaert, Jim Spohrer, Mark Steedman, Claes Strannegård, Neset Tan, Tadahiro Taniguchi, Moshe Vardi, Karina Vold, Michael Witbrock, Michael Wooldridge, Hiroshi Yamakawa.

Secondly, we wish to thank our colleagues in the Centre for Educational Research and Innovation (CERI). Tia Loukkola, Head of CERI, provided oversight, direction and valuable advice during the process. Colleagues from the Programme for International Assessment of Adult Competencies (PIAAC) and the Programme for International Student Assessment (PISA) made important contributions to the analysis. Colleagues within the Directorate for Education and Skills communications team and the Public Affairs and Communications Directorate contributed to both formatting and the preparation of the publication.

Our thanks are extended to Mark Foss, who made substantive and structural editing to the publication, ensuring for coherent, comprehensible reading.

We are grateful for the encouragement and support of the CERI Governing Board in the development of the project.

This publication contributes to the OECD's Artificial Intelligence in Work, Innovation, Productivity and Skills (AI-WIPS) programme, which provides policy makers with new evidence and analysis to keep abreast of the fast-evolving changes in AI capabilities and diffusion, and their implications for the world of work. The programme aims to help ensure that adoption of AI in the world of work is effective, beneficial to all, people-centred and accepted by the population at large. AI-WIPS is supported by the German Federal Ministry of Labour and Social Affairs (BMAS) and will complement the work of the German AI Observatory in the Ministry's Policy Lab Digital, Work & Society. For more information, visit <https://oecd.ai/workinnovation-productivity-skills> and <https://denkfabrik-bmas.de/>.

Table of contents

Foreword	3
Acknowledgements	4
Executive summary	9
1 Overview	12
Overview of the AI and the Future of Skills project	14
Lessons learnt from the first project stage	16
The second stage of the project	17
Outline of the structure of the report	20
References	22
Notes	23
2 Eliciting expert knowledge: Methods and challenges	24
Methods for eliciting expert judgement	26
Large-scale experiment: How many experts can be engaged and through what incentives?	30
Task framing used to collect expert judgement on AI capabilities	33
Establishing consensus: Quantitative disagreement versus qualitative agreement	35
Conclusions: Challenges and future directions	37
References	38
Notes	39
3 Assessing AI capabilities with education tests	40
Rationale for assessing AI capabilities with education tests	42
Overview of the education tests used	43
Methodology for collecting expert judgement on AI with education tests	47
Results	51
Lessons learnt	58
The way forward	60
References	61
Annex 3.A. Analyses of the PIAAC and PISA studies using an alternative approach	63
Notes	64
4 Occupational tests	65
Rationale for collecting expert judgement on AI with complex occupational tasks	66
Occupational tasks from certification and licensure examinations	67
Selection of occupations and examination tasks	70

The way forward	73
References	75
Notes	77
5 Assessing AI capabilities on occupational tests	78
Collecting expert judgement on performance tests of occupational tasks	79
Evaluation of AI and robotics capabilities on tasks and subtasks	81
Evaluation of AI and robotics capabilities on capability scales	90
The way forward	94
References	95
Annex 5.A. Categories of AI capabilities	96
Notes	98
6 A framework for characterising evaluation instruments of AI performance	99
Characterising AI evaluation instruments	100
Evaluation instrument selection and rating methodology	103
Analysis of rater consistency	104
Analysis of facet values	105
Conclusion	110
References	112
Annex 6.A. Supplementary tables	116
Notes	118
7 AI direct tests: LNE and NIST evaluations	119
The need for systematising AI and robotics evaluations	120
Framework structure	121
Evaluation campaigns of AI capabilities	124
Limitations and uncovered tasks from AI evaluations	138
Conclusion	139
References	140
Annex 7.A. Low functionality levels AI tasks of evaluation campaigns across the three major fields of NLP: computer vision and robotics	142
Annex 7.B. Detailed facet characteristics attributions of the LNE and NIST evaluations	145
Notes	145
8 Towards a synthesis of language capability in humans and AI	146
Benchmark tasks: Narrow versus strong AI	147
Conceptual framework of language competences	148
Mapping major language benchmarks to the human language competence framework	150
Language understanding: AI vs. human	151
Language generation: AI vs. Human	157
Update of AI language competences post ChatGPT release	160
Conclusion	162
References	163
Annex 8.A. Natural Language Processing research areas	165
Notes	166
9 Project goals, constraints and next steps	167
Potential sources of information about AI capabilities	168
Information needed about AI's implications for education and work	172

Next steps for the project	174
References	177
Notes	177

FIGURES

Figure 1.1. Sources of AI assessments	16
Figure 2.1. The effect of incentives on the final response rate	32
Figure 2.2. Self-reported motivation to complete the survey	32
Figure 3.1. PIAAC Literacy and Numeracy – Sample items	45
Figure 3.2. PISA Science – Sample item	46
Figure 3.3. AI literacy performance in 2016 and 2021, by question difficulty	52
Figure 3.4. AI numeracy performance in 2016 and 2021, by question difficulty	53
Figure 3.5. Predicted AI performance on PISA science questions in 2022 by core experts and larger expert group, by question difficulty	54
Figure 3.6. Literacy performance of AI and adults of different proficiency	55
Figure 3.7. Divergence in experts' evaluations in different assessments	56
Figure 3.8. Share of questions that receive more than 20% of uncertain ratings in different assessments	57
Figure 3.9. Experts' ratings of AI and GPT-3.5 performance on PISA science questions	58
Figure 5.1. AI and robotics performance on entire task, by task format	82
Figure 5.2. Distribution of expert ratings of AI and robotics performance on entire task	83
Figure 5.3. Average AI and robotics performance, by expert and expertise	84
Figure 5.4. AI and robotics performance in broad capability domains, by task and expertise	85
Figure 5.5. AI and robotics performance on subtasks, by complexity level and broad capability domain, mid-2022	86
Figure 5.6. Expert descriptors of complexity levels of broad capability domains	87
Figure 5.7. AI capability expert ratings and their comparison to the ratings of the first study	91
Figure 5.8. AI capability expert ratings, by task	92
Figure 6.1. Rater agreement across all facets	105
Figure 6.2. Rater value selection on validity facets	106
Figure 6.3. Raters value selection on consistency facets	108
Figure 6.4. Raters value selection on fairness facets	109
Figure 7.1. Rater values selection on validity facets for eight evaluation campaigns by NIST and LNE	125
Figure 7.2. Rater values selection on consistency facets for eight evaluation campaigns by NIST and LNE	126
Figure 7.3. Rater values selection on fairness facets for eight evaluation campaigns by NIST and LNE	127
Figure 8.1. General relationship between human language and NLP difficulty levels with respect to the input and output format moving from text to speech and vice versa	149
Figure 8.2. Relationship between required minimum human language competence level and state-of-the-art NLP system performance for a sample of NLP tasks	150
Figure 8.3. Example dialogue from the TRAINS corpus	152
Figure 8.4. Sample constituency parse tree of English sentence	155
Figure 9.1. Conceptual scale reflecting AI performance levels	175

TABLES

Table 2.1. Major EKE protocols	27
Table 2.2. Methods used to collect expert judgement in the AIFS project	29
Table 2.3. Response rate	31
Table 2.4. Task framing and response format in the AIFS project	35
Table 4.1. Selected occupations	70
Table 4.2. Selected occupational tasks	72
Table 6.1. Primary testing domain of sampled evaluation instruments	103
Table 6.2. Type of sampled evaluation instruments	104
Table 7.1. Text processing and comprehension high-level task examples and associated evaluation campaigns	129
Table 7.2. Speech processing high-level task examples and associated evaluation campaigns	130

Table 7.3. Recognition high-level task example and associated evaluation campaigns	131
Table 7.4. Motion analysis high-level task examples and associated evaluation campaigns	132
Table 7.5. Locomotion high-level task examples and associated evaluation campaigns	134
Table 7.6. Manipulation task examples and associated evaluation campaigns	136
Table 8.1. NLP research areas with type and level of language competence required for humans	149
Table 8.2. Datasets in the GLUE benchmark	153

Annex Table 3.A.1. List of online figures for Chapter 3	63
Annex Table 5.A.1. Categories of AI capabilities	96
Annex Table 6.A.1. Overview of Evaluation Instruments	116
Annex Table 7.A.1. Low functionality level tasks of evaluation campaigns associated with the NLP field	142
Annex Table 7.A.2. Low functionality level tasks of evaluation campaigns associated with the Computer Vision field	143
Annex Table 7.A.3. Low functionality level tasks of evaluation campaigns associated with the Robotics field	144
Annex Table 8.A.1. Natural Language Processing research areas with at least one benchmark task	165

BOXES

Box 1.1. Types of AI measures discussed in the report	19
Box 2.1. Evolution of assessment instructions in the AIFS project	34
Box 3.1. Use of education tests in AI evaluation	43
Box 3.2. Example items from PIAAC and PISA	45
Box 4.1. Direct assessment of large language models on written professional certification tests	71
Box 7.1. Facet characteristics of the LNE and NIST evaluations vs. those of benchmark tests	124
Box 8.1. Transformer models	148
Box 8.2. Example of anaphora and coreference resolution	152
Box 8.3. GLUE benchmark	153

Follow OECD Publications on:



<https://twitter.com/OECD>



<https://www.facebook.com/theOECD>



<https://www.linkedin.com/company/organisation-eco-cooperation-development-organisation-cooperation-developpement-eco/>



<https://www.youtube.com/user/OECDiLibrary>



<https://www.oecd.org/newsletters/>

This book has...

StatLinks

A service that delivers Excel® files from the printed page!

Look for the *StatLink* at the bottom of the tables or graphs in this book. To download the matching Excel® spreadsheet, just type the link into your Internet browser or click on the link from the digital version.

Executive summary

As artificial intelligence (AI) and robotics technologies continue to expand their scope of applications across the economy, understanding their impact becomes increasingly critical.

The AI and the Future of Skills (AIFS) project at OECD's Centre for Education Research and Innovation (CERI) is developing a comprehensive framework for regularly measuring AI capabilities and comparing them to human skills. The capability measures will encompass a wide range of skills crucial in the workplace and cultivated within education systems. They will establish a common foundation for policy discussions about AI's potential effects on education and work.

The AIFS project has undergone two phases of developing the methodology of the assessment framework. The first phase focused on identifying relevant AI capabilities and existing tests to evaluate them. It drew from a wealth of skill taxonomies and assessments across various disciplines, including computer science, psychology and education.

The second phase, the focus of this report, delves deeper into methodological development. It comprises three distinct exploratory efforts:

Rating AI on education tests using expert judgement

Education tests offer a valuable means of comparing AI to human capabilities in domains relevant to education and work. The project carried out two studies to explore the use of education tests for collecting expert judgements on AI capabilities. The first study, conducted in 2021/22, followed up an earlier pilot study, asking experts to evaluate AI's performance on the literacy and numeracy tests of the OECD's Survey of Adult Skills (PIAAC). The second study collected expert judgements of whether AI can solve science questions from the OECD's Programme for International Student Assessment (PISA).

Purpose

The studies aimed to refine the assessment framework for eliciting expert knowledge on AI using education tests. They explored different test tasks, response formats and rating instructions, along with two distinct assessment approaches: a "behavioural approach" used in the PIAAC studies, drawing on smaller expert groups engaging in discussions, and a "mathematical approach" adopted in the PISA study, relying more heavily on quantitative data from a larger expert pool.

Lessons learnt

This work showed that there are limits to obtaining robust measures of AI capabilities by surveying experts. Especially in domains that are not the centre of current research, consensus evaluations are hard to reach. In addition, recruiting and engaging experienced experts is costly.

Rating AI on occupational tests

Two exploratory studies extended the rating of AI capabilities to tests used to certify workers for occupations. These tests present complex practical tasks typical in occupations, such as a nurse moving a paralysed patient, or a product designer creating a design for a new container lid. Such tasks are potentially useful as a way of providing insight into the application of AI techniques in the workplace.

Purpose

The inherent complexity of occupational tasks makes them different from the questions contained in education tests. 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. The two studies explored the use of different survey instruments and instructions for collecting reliable and valid expert evaluations on these tasks.

Lessons learnt

Rating AI performance on occupational tasks proved challenging. The rating difficulty related to predicting contextual factors that can potentially affect AI performance and to specifying the underlying capability requirements for each task. On the other hand, the studies suggested the possible use of occupational tasks for better anticipating how occupations might evolve as new AI capabilities emerge. As a result, occupational tasks will be used to understand the implications of AI for work and education rather than for gathering expert judgements on AI capabilities.

Direct measures of AI capabilities

Recognising the limitations of expert judgement, the project initiated an exploration of measures derived from direct evaluations of AI systems. These benchmark tests offer a diverse range of evaluations but vary in quality, complexity, and target capabilities. To navigate this landscape, the project commissioned experts to explore their uses for the project.

Purpose

The project needed to find ways to select good-quality benchmarks, categorise them according to AI capabilities and systematise them into single measures. It commissioned experts to work on each of these tasks. Anthony Cohn and José Hernández-Orallo developed a method for describing the characteristics of benchmark tests to guide the selection of existing measures for the project. Guillaume Avrin, Swen Ribeiro and Elena Messina presented evaluation campaigns of AI and robotics and proposed an approach for systematising them according to AI capabilities. Yvette Graham reviewed major benchmark tests in the domain of natural language processing and developed an integrated measure based on the reviewed tests.

Lessons learnt

Using direct measures to develop valid indicators of AI capabilities is a challenging but promising direction because of the large number and variety of direct measures available. The measures evolve rapidly as the field itself, which requires an approach to synthesising them conceptually. In addition, the measures often omit a comparison to human capabilities, which requires additional steps to add this reference. The preliminary work on direct measures suggests ways of addressing these two challenges.

Future directions

Both expert judgements and direct AI measures are necessary to develop indicators of AI capabilities that are understandable, comprehensive, repeatable and policy relevant. The project's third phase is working on a concrete approach for developing such indicators in different domains. This approach draws on experts to review, select and synthesise direct AI measures into a set of integrated AI indicators. These will be complemented with measures obtained from expert evaluations in areas where direct AI assessments are lacking. The resulting AI indicators will then be linked to measures of human competences and examples of occupational tasks to derive implications for education and work. They should aid decision-makers in determining necessary policy interventions as AI continues to advance.

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 robotics¹ 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]).

2 Eliciting expert knowledge: Methods and challenges

Abel Baret, Nóra Révai, OECD

Gene Rowe, Gene Rowe Evaluations

Fergus Bolger, Anglia Ruskin University

This chapter delves into the methodology of collecting expert judgement in the AI and the Future of Skills project. It provides an overview of the project's journey in refining its methodology and discusses associated challenges and considerations. The chapter begins by exploring the different methods of expert knowledge elicitation based on the research literature and discusses their relevance to the project. It then addresses key questions such as the number of experts required for reliable assessments, the framing of tasks for experts, and the aggregation and interpretation of expert judgements. The chapter concludes by offering points of consideration for the project's long-term trajectory.

This chapter reports on the journey of refining the project's methodology to collect expert judgement on the capabilities of artificial intelligence (AI). As its main approach to developing measures of AI capabilities, the AI and the Future of Skills (AIFS) project initially relied on the judgement of computer scientists to assess these capabilities based on questions in human tests. This idea originates in a need to support the policy community in planning education and employment policies with a sound knowledge of the progress in AI capabilities and how that compares to human skills.

The study chose to focus initially on human tests rather than on direct measures for several reasons. There are many direct measures of AI system performance (benchmarks, competitions, formal evaluation campaigns). However, these often measure performance on specific, narrow tasks. In addition, these are not synthesised into broader capability areas that would be meaningful for policy makers. These direct measures also miss certain skills that are important for humans and do not always allow for a comparison between machine and human performance. Therefore, the OECD decided to develop measures reflecting computer scientists' judgements using human tests as a first approach.

This approach requires establishing a robust methodology for collecting expert judgements that is valid and reliable, and ideally reflects a consensus of the expert community. Such a methodology involves recruiting and engaging the right experts, a well-established process for collecting expert judgement, a well-framed task for experts, an instrument (test questions or tasks) that allows computer scientists to assess AI capabilities correctly and a method that yields the consensual result of experts' judgements.

The precursor of the project was a pilot study in 2016 that asked computer scientists whether AI technology of the time and five years from then could answer the questions in the OECD's Survey of Adult Skills (in the Programme for International Assessment of Adult Competencies [PIAAC]) (Elliott, 2017^[1]). The pilot revealed several strengths but also some weaknesses in its methodology.

The project thus set out to consolidate its method to collect expert judgements in two main ways. First, it explored the literature on expert knowledge elicitation (EKE) and sought advice from experts in this methodological field. A meeting in March 2021 brought together experts to discuss the various methods for EKE and assess their relevance and feasibility for the AIFS project. Second, it conducted a series of exploratory studies in which the project tried out different methods to answer the following questions:

- Which EKE method is the most suitable to collect expert judgement on AI capabilities using human tests?
- How many experts are needed to obtain a reliable assessment of AI capabilities? How can they be identified, recruited and engaged?
- How does the task need to be framed so that experts have a unified understanding of the task and are able to provide a precise judgement of AI capabilities?
- How can we aggregate and interpret the results of expert judgement to obtain a single measure of AI capabilities?

In March 2022, the project held an expert meeting that discussed aspects of the methodology of collecting expert judgement. This included the overall framing of the task and the nature of information needed about the test used, as well as the specific instructions and response formats.

This chapter gives an overview of the different aspects of EKE based on the literature and discusses their application in the context of the project. It describes the evolution of methodologies across the exploratory studies along three main factors: the method of collecting expert judgements, the number of experts involved in the assessment and the framing of tasks for experts. The chapter then discusses the level of consensus in expert judgements on AI capabilities. It concludes with summarising the major developments and challenges in the methodology, offering a few points of consideration for the longer term. The subsequent chapters give details about the methods the project used in the series of exploratory studies.

Methods for eliciting expert judgement

Eliciting expert judgement has its own methodological literature referred to as EKE (O’Hagan et al., 2006^[2]) or Structured Expert Judgement (Cooke, 1991^[3]; Hanea et al., 2021^[4]). This area emerged from the necessity to supplement sparse or missing empirical, scientific evidence with expert judgement that can serve as the basis for decisions and policy making. EKE – defined as structured group techniques for the elicitation of judgements of uncertain quantities – is a relatively new area. However, it is based on earlier techniques for surveying and eliciting expert knowledge and group techniques [e.g. (Rowe, 1992^[5]; von der Gracht, 2012^[6]; Kahneman, Slovic and Tversky, 1982^[7]; Linstone and Turoff, 1976^[8])]. In the case of AI capabilities, the motivation to use expert judgement is due to the scattered and unstructured nature of available direct assessments, and their unsuitability for the policy community (see details in Chapter 9).

Behavioural and mathematical approaches to eliciting expert knowledge

EKE methods attempt to elicit judgements from experts that are as reliable and as valid as possible. This involves aggregating across several different opinions in a carefully managed process that helps reduce individual bias (e.g. resulting from beliefs and cognitive or social dispositions rather than scientific findings) and possible distortions resulting from group interactions. Further, quantitative judgements carry varying degrees of uncertainty, which are important to capture when informing policy decisions. In addition, EKE can elicit qualitative judgements from experts, either as input to decision making in their own right or to support quantitative judgements, for example, as rationales for them.

To inform decision making, a summary of judgements by groups of experts into a single estimate (or perhaps two or three if there are distinct schools of thought) is more useful than numerous individual judgements. Aggregation across multiple judgements also serves to reduce random error in those judgements. EKE techniques can use behavioural and mathematical aggregation of judgements or a mixture of the two to arrive at a single group judgement (O’Hagan et al., 2006^[2]).

- Behavioural aggregation involves interacting experts – facilitated or otherwise – coming to a consensus.
- Mathematical aggregation means averaging over different individual judgements. This can be done with equal weights given to each expert, or different weights (e.g. “performance weights” based on an assessment of individual expert ability).

The main difference between the two approaches is the degree of interaction between experts. In behavioural approaches, there is usually a high level of interaction among experts (either in a facilitated discussion or freely in a meeting, by e-mail or otherwise). In purely mathematical approaches, experts do not interact with each other (Rowe, 1992^[5]).

Behavioural aggregation can be applied to both qualitative and quantitative judgements and tolerates different schools of thought. If well-managed, it can allow experts to weight themselves in terms of their respective knowledge of an issue (e.g. by moving towards the positions of those with more expertise). However, if not well-managed, the process of behavioural aggregation can lead to biased outcomes resulting from social and cognitive biases such as group polarisation, overconfidence and groupthink (Lichtenstein, Fischhoff and Phillips, 1982^[9]; Myers and Lamm, 1976^[10]; Turner and Pratkanis, 1998^[11]). Mathematical aggregation with equal weights is simple but does not consider individual differences in expertise. Performance weighting has the advantage to account for such differences (Hanea et al., 2021^[4]). However, it has practical difficulties such as obtaining valid performance weights and may risk alienating the experts (Bolger and Rowe, 2015^[12]; Bolger and Rowe, 2015^[13]).

Behavioural and mathematical aggregation represent the two extremes in EKE. In between, other approaches combine both behavioural and mathematical elements. The main steps of different EKE

approaches are commonly recorded as protocols. Table 2.1 describes the major protocols that have been developed to collect expert judgement.

The protocols differ in the degree they use behavioural and mathematical aggregation and thus require varying degrees of interaction among experts. They also differ in the extent and nature of facilitation needed. Facilitators' skills can be key to the successful organisation and running of interactive group processes. They involve the ability to carefully guide discussions to include every issue intended for debate, avoid inserting their own viewpoints and ensure that discussions are not prematurely closed off. Facilitators also need to be able to involve all experts equitably and use continual summarising processes to confirm that all points are accurately understood and collated.

Table 2.1. Major EKE protocols

Group EKE protocol (and Reference)	Description	Aggregation type (MA: mathematical BA: behavioural aggregation)
One-shot surveys	A questionnaire for experts to complete individually, with responses usually averaged (with equal weighting) to indicate group judgement (and distributions used to indicate response variability).	MA
Classical method (CM) (Cooke, 1991 ^[3])	Experts are usually "tested" individually and then their judgements are combined mathematically and unequally according to performance weights based on testing results	MA
Delphi method (Linstone and Turoff, 1976 ^[8] ; Rowe, Wright and Bolger, 1991 ^[14])	Experts complete a survey anonymously and individually, receive the (summarised) responses from a facilitator and revise their responses. This can be repeated in further rounds. Delphi methods vary according to how they are precisely operationalised (e.g. Classical, Policy, Real-Time). Well-suited to online delivery	MA with equal weighting and varying degrees of BA.
Investigate-Discuss-Estimate-Aggregate (IDEA) (Hemming et al., 2018 ^[15])	As in CM, experts first individually make judgements of "seed questions" – for performance weighting – and the target questions. Next there is a (usually online) meeting of all experts with a facilitator to discuss the initial estimates and ensure a common understanding of the judgement task. Finally, the experts make judgements of target questions individually again, which are aggregated using the performance weights.	MA and some BA (although discussion primarily meant for problem clarification).
Nominal Group Technique (NGT) (Delbecq and Van de Ven, 1971 ^[16] ; Delbecq, Van de Ven and Gustafson, 1975 ^[17])	A facilitated group approach that allows face-to-face discussion with individual and anonymised estimations of the solution before and after discussion, and equal-weighted judgement of the final (post-discussion) estimates	BA and MA with equal weighting.
Facilitated group processes (e.g. Sheffield method) (Gosling, 2018 ^[18])	Interactive group processes that are generally held face-to-face (although real-time online processes are also possible). They rely on careful facilitation to ensure focused discussion and equal participant contribution, with the aim being group consensus.	BA

When determining which protocol to choose for a particular application, a number of factors need to be considered.

- Number of experts: behavioural methods are suitable for a small number of experts; mathematical approaches allow for collecting judgement from a large number of experts.
- Range of experts: heterogeneous expert groups (e.g. in terms of disciplinary background) favour a behavioural method where a facilitator can help overcome differences for example in knowledge base and language.
- Number of questions: mathematical aggregation is easier where there are a large number of questions.
- Nature (complexity) of questions: behavioural method is more suitable for complex tasks/questions that require substantive input to ensure a common understanding and more in-depth discussions.
- Nature of response: mathematical methods require quantitative response options sometimes complemented with qualitative responses (e.g. rationales); behavioural methods can be suitable for both quantitative and qualitative responses.

Additional considerations for the method include its cost, feasibility of recruiting and engaging experts, and feasibility of achieving consensus or establishing a single aggregate measure.

EKE methods used in the AIFS project

EKE in the AIFS project involves asking experts about whether AI can answer specific questions or carry out specific tasks. The pilot study, which used the OECD's PIAAC survey, opted for a facilitated face-to-face group discussion over two days. Such an extensive, in-depth discussion was necessary to elicit a feasible and meaningful framing of the rating task for experts, to identify difficulties and agree on the overall approach. However, this method has its trade-offs: it is expensive (travel and accommodation costs for all experts); it limits the number of experts able to participate in the exercise; it is time-consuming without much flexibility (experts cannot choose the best time for themselves to go through the 113 questions of the PIAAC survey); and it leaves little room and time for individual reflection.

For the more recent exploratory studies – an update using the OECD's PIAAC survey, a study using the OECD's Programme for International Student Assessment (PISA) test (see Chapter 3 for details), and studies using selected occupational tests and tasks (see Chapters 4 and 5) – the project tested other methods. In choosing the methods, the project considered the following:

- The selected human tests often involve different areas of computer science, such as natural language processing (NLP), computer vision and robotics. In addition, expertise in other disciplines, such as organisational and industrial psychology, is required to clarify what a test or task involves on the human side. Therefore, the expert group is relatively heterogeneous.
- Unlike many of the EKE tasks reported in the literature, this is not primarily a forecasting task. The capabilities of current technology are only available in a highly technical language, they are scattered, not systematised and evolve rapidly. A high level of expertise is necessary to be aware of and understand current AI capabilities. Projections for the future are generally based on ongoing research grants, which again requires expert knowledge and involvement in research and development.
- Some studies include many questions (PIAAC and PISA tests) and some are complex in nature (occupational tests). Most questions require expertise in several subdomains of computer science (e.g. computer vision and NLP).
- It is important to test whether using a small number of experts as opposed to a large number yields substantively different results.
- It is important to test the feasibility of different approaches in terms of costs, human resources, expert recruitment, etc.
- Reaching consensus is highly desirable given that the task is to gauge current computer capabilities, which should be knowable. While consensus among experts would also facilitate informing the policy community and drawing policy implications, it is vital to draw their attention to existing debates (dissensus) within the computer science community if these exist.

The COVID-19 pandemic (2020-21) prevented the project from organising face-to-face meetings, but made online meetings easier with improved platforms and people getting used to them.

Based on the above considerations, the project opted for testing a combination of mathematical and behavioural methods to elicit experts' judgement on AI capabilities. Table 2.2 summarises the methods used, and the number and background of experts involved.

Table 2.2. Methods used to collect expert judgement in the AIFS project

	EKE method	Experts
PIAAC 2016	Facilitated group discussion: <ul style="list-style-type: none"> • 2 days • In-person 	<ul style="list-style-type: none"> • N=11 • Computer scientists
PIAAC 2021 follow-up 1	Modified Delphi method: <ul style="list-style-type: none"> • Online survey (round 1) • Online group meeting • Online survey (round 2)* 	<ul style="list-style-type: none"> • N=11 • Computer scientists, Cognitive and I/O psychologists
PIAAC 2022 follow-up 2	Modified Delphi method: <ul style="list-style-type: none"> • Online survey (round 1) • Online group meeting • Online survey (round 2)* 	<ul style="list-style-type: none"> • N=4 • Computer scientists
PISA 2022 core experts	Modified Delphi method: <ul style="list-style-type: none"> • Online survey • Online group meeting 	<ul style="list-style-type: none"> • N=12 • Computer scientists, Cognitive and I/O psychologists
PISA 2022 new experts	Online survey	<ul style="list-style-type: none"> • N=170 invited • R=33 respondents • Computer scientists
Occupational 2022	Modified Delphi method: <ul style="list-style-type: none"> • Online survey • Online group meeting 	<ul style="list-style-type: none"> • N=12 • Computer scientists, I/O psychologists

Note: *Completing the same survey again to modify initial judgements was offered to experts, but none of them actually did it. This option was thus dropped from subsequent studies.

With respect to working with a heterogeneous expert group on complex test questions, the project found a **modified Delphi method** as the most appropriate approach for most of the assessments.

Delphi is a structured group technique that consists of at least two rounds of surveys collecting experts' ratings, with feedback on the ratings provided between rounds. The iteration of survey rounds continues until consensus among experts is reached. During each round, experts provide their ratings anonymously and independently from each other. This should reduce potential bias from social conformity or from dominant individuals who impose their opinions on the group. By contrast, the feedback provided after each round should enable social learning and the modification of prior judgements due to new information. This feedback should ultimately increase consensus between experts.

Designing the appropriate method needs to take into account the features of the task of assessing current AI systems' capabilities (described above). Importantly, a range of specialised knowledge is required in sub-fields of AI. For some tests (e.g. PIAAC literacy and PISA reading), all experts are generally aware of the current state of the art in relevant AI domains. Other tests, such as rating occupational tasks, require more specialised knowledge (e.g. in robotics). In either case, individual experts cannot possibly know all existing AI applications, recent research results or other details that may be relevant for the evaluation. For example, only one or a few experts may have knowledge on particular AI systems that can perform a task. To facilitate consensus, experts should be able to communicate such information to the group at any point of the rating process.

For this reason, in contrast to a classical Delphi approach, a high degree of interaction among experts is more suitable for assessing AI capabilities on tests/tasks. Thus, after the first round, the project organised a three-hour online meeting in all exploratory studies. This meeting allowed experts to discuss the feedback they received on the survey results, exchange ideas and share references to recent research results. After the meeting, experts were invited to revise their judgements provided in the survey based on the group discussion.

Overall, the experts and the project team were satisfied with the modified Delphi method in at least two regards. Experts appreciated exchanging references and discussing ideas. Meanwhile, the project team could elicit the group's overall assessment of AI capabilities. This was true even if there was disagreement in the ratings experts provided through the survey.

However, one key feature of the Delphi method did not prove feasible. Although the project team asked experts to revise their responses if their views have changed, they did not go back to the survey to modify their ratings after the meeting. This may be because the interactions during the meeting provided an opportunity for experts to explain their judgements and reconsider them in light of a better understanding of the scope of the task. On numerous occasions, they expressed how they would modify their judgement with the new understanding at the meeting. The experts had numerous test questions to review, and it took time to provide judgements on each of them. Consequently, it was generally not practical to push them to do more than a one-time survey and one meeting given that they often provided a modified judgement at the meeting already. As a result, the quantitative (mathematical) aggregation of expert judgements needed to be complemented by the qualitative aggregation resulting from the meeting (see Chapters 3 and 5 for more details).

Large-scale experiment: How many experts can be engaged and through what incentives?

The pilot study with PIAAC, as well as its follow-up, relied on a core group of 10-15 experts who have worked closely with the project team from the outset. To test the feasibility of involving substantially more experts and if this would yield different and/or more robust results, the project conducted a large-scale version with a different assessment – the PISA science assessment (see Chapter 3). Having a large sample of computer scientists willing to invest substantially in providing judgements was expected to be challenging. The team thus tested different strategies in approaching and engaging experts. The goal of the experiment was to answer the following questions:

- How many experts can be identified and contacted within a limited timeframe?
- What response rate can be expected?
- Is an incentive necessary to engage experts? And if so, which one is the most effective?

Recruiting and engaging experts: Outreach and incentives

The first challenge was to identify a large number of experts with the appropriate background. The list of experts had to cover all relevant domains of computer science (e.g. NLP, computer vision, reasoning) and demonstrate diversity, in particular with respect to gender and geographical coverage. To compile the list, the project used snowballing, starting with recommendations from its already engaged small group of experts (henceforth “core experts”). In addition, the team identified and scanned the webpages of relevant research laboratories, conference attendee lists, and public and private organisations. Some 170 experts were selected (of whom 119 were recommended by our core experts) and contacted. In addition, the project reached out to 19 graduate students. Overall, the final list included 111 males and 59 females and covered 19 countries.

The second challenge was to convince the experts to participate in the study, i.e. to try to achieve a high response rate. The project tested different incentives to determine the most effective way to engage experts. All graduate students were offered a EUR 250 honorarium. Meanwhile, the experienced computer scientists were randomly distributed in four groups of 11 participants that had different incentives to complete the survey:

1. *Honorarium group*: receiving an EUR 800 honorarium;

2. *Co-authorship group*: offered to be co-authors of a future report;
3. *Honorarium + Co-authorship group*: receiving both incentives;
4. *No incentive group*¹.

To reach out to experts, the project used a foot-in-the-door technique. As such, it drew on evidence that people are more likely to agree to a request when they have already made a commitment to a similar action (Freedman and Fraser, 1966^[19]). The first e-mail briefly presented the project, and issued invitations to participate in a survey (without giving many details) and to join the project community. The e-mail also mentioned the name of the core expert who recommended them, when applicable. It asked about experts' interest to learn more about the project and the survey. To experts recommended by the project's core experts, the first e-mail also offered an online call to discuss the project. Experts who answered the first e-mail and expressed interest received a second e-mail with detailed information about the survey and their respective incentives.

Results: Response rate and effects of incentives

Table 2.3 shows the response rates. A quarter of the targeted experts showed initial interest (i.e. responded to the first e-mail); 77% of these respondents were experts recommended by the project's core experts. This shows the importance of snowballing and referrals. Slightly less than the half of experts who showed initial interest actually completed the survey. Most experts who did not complete the survey informed us of their withdrawal, and typically referred to lack of time or interest in the survey. Among them, 15 nonetheless expressed interest in meeting with the team to learn about the project.

Table 2.3. Response rate

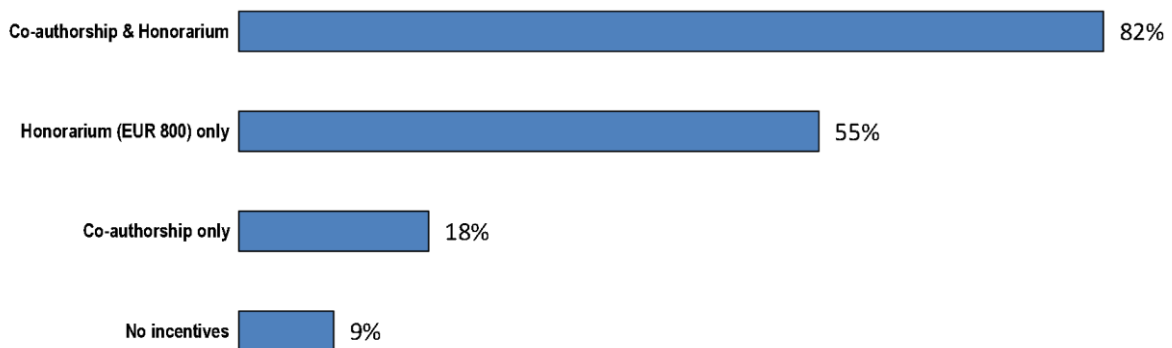
	Experienced experts		Graduate students	Total
	Recommended by core experts	Identified		
Total number targeted	119	51	NA*	189+
Answered the first e-mail (as percent of those targeted)	34 (28.6%)	10 (19.6%)	19	63
Filled in the survey	13	5	15	33
Final response rate (with respect to those who answered the first e-mail)	38%	50%	79%	52%


Note 1: *Graduate students were initially reached via a form sent to university forums and the project's core experts.

Note 2: Statistical tests (chi-square) did not yield a significantly different response rate between the group recommended by experts and those identified by the team.

Unsurprisingly, the combination of incentives had the strongest effect on completion, while the group receiving no incentives had a low response rate (Figure 2.1). Clearly, money matters most. More than half of participants from the Honorarium group completed the survey, as opposed to less than one in five among those offered co-authorship.

Figure 2.1. The effect of incentives on the final response rate



StatLink  <https://stat.link/g5k6i0>

At the end of the survey, experts were asked about their motivation to answer via a multiple choice of four options (Figure 2.2). The “*interest in the nature of the assessment and the test items*” appeared to be the strongest self-reported motivation factor, whereas *co-authorship opportunity* was the least important self-reported factor.

Figure 2.2. Self-reported motivation to complete the survey



StatLink  <https://stat.link/chife0>

In sum, the experiment highlighted several challenges of engaging a large number of experts in such a time-consuming activity. First, it is difficult to identify many experts with such specific expertise. Crowdsourcing experts would not ensure the level of expertise needed for the task. Second, it is difficult to engage them in completing a long and complex survey. Offering money (if possible, together with co-authorship) may ensure acceptable response rates. However, this is obviously a costly measure, especially with large samples.

Computer scientists with the high level of expertise required for this task are generally very busy. In addition, those working in industry often do not feel comfortable participating in such an exercise because they are bound by business secrecy. Moreover, the financial incentives might need to be larger to be effective for this group.

Survey length could in principle be reduced using incomplete block design, i.e. each respondent only answers a smaller subset of the questions. However, this method requires a large sample of experts to ensure that missing values can be reliably estimated. Many questions require several domains of expertise at the same time (e.g. they include a visual element and require language understanding necessitating

expertise in both computer vision and NLP). Therefore, sorting questions based on subdomain expertise is not possible.

Conducting a large-scale survey repeatedly to collect expert judgement in the domain of AI capabilities is therefore highly challenging, if feasible at all.

Task framing used to collect expert judgement on AI capabilities

This section discusses the challenge of framing the rating task for experts in a way that ensures common understanding and reliable assessment of AI capabilities.

Task framing and instructions

The pilot study that used the PIAAC test asked experts to give a rating (Yes, No or Maybe) of AI systems' ability to solve each test question after one year development and a cost limit of USD 1 million. The latter parameters were defined in the meeting of the pilot study (Elliott, 2017^[1]) to specify what it means to rate current technology even if no off-the-shelf system is available.

The same instruction was kept for the first exploratory study that updated the PIAAC pilot in 2021. However, some limitations of this instruction emerged. These included experts interpreting the scope of abilities covered in the assessments differently. For instance, some experts focused on AI systems' narrow ability to answer the given set of questions, while others imagined that the questions were representative of a broad underlying capability (see Chapter 3 and OECD (2023^[20])). Some also judged the USD 1 million parameter as unrealistic with regard to commercial AI development projects in the field. These limitations suggested the need for developing a finer framing for the assessments.

To address the above concerns, the project team created a **framework document** for the subsequent exploratory studies that gives more details on the assessments and describes the characteristics of the test questions:

- what the test measures in terms of human skills (e.g. literacy skills)
- how the test measures this skill (e.g. multiple choice questions about simple comprehension of a text)
- factors affecting question difficulty (e.g. interpretation required)
- scoring rubrics used to evaluate test takers' performance.

The framework document also included examples of test items to give experts a sense of what to expect from the assessment. The examples can be considered a representative set of training data that define the scope of abilities. This helps experts imagine a machine learning system that could be developed.

Instructions were also changed to account for the problems mentioned above (see Box 2.1 and Table 2.4 for the evolution of instructions and task framing). In particular, the description of “current computer techniques” changed and the scope of abilities was specified through examples. The prompt to imagine an AI system that answers the questions clarified that we need one integrated system as opposed to fine-tuned systems for each question.

Box 2.1. Evolution of assessment instructions in the AIFS project

Extract from PIAAC 2016 and 2021 assessment instruction

[...] You will be asked to evaluate the capacity of AI technologies to correctly answer the PIAAC questions. In making your judgement, please consider the following:

- Please consider “current” computer techniques, meaning any available techniques that have been addressed in the literature that we can describe their capabilities and limitations.
- Please consider techniques that might need “reasonable advance preparation”, [...] thinking about a development team receiving detailed information about the types of questions included in the test and being given one year and USD 1 million funding to build and refine a system to work with such questions using current techniques.

Extract from PISA 2022 and PIAAC follow-up assessment instruction – Imagined AI system

[...] The questions are presented in different formats (including pictures, texts and numbers) and are designed to resemble real-life tasks in work and personal life. You will be asked to:

- briefly describe a high-level approach for an AI system built to answer the questions on the PISA science assessment
- evaluate the likely performance of that AI system on different questions from PISA.

To help you understand the domain, a document describing the framework for the PISA science test and providing a set of ten example questions was provided to you beforehand. [...]

In designing the high-level approach for your AI system, you had to consider any “current” computer techniques [...]. The point is that the design for your imagined AI system should involve the application of existing AI techniques, not research to develop new approaches.

Extract from Occupational tasks assessment instruction

[...] You will be asked to evaluate the capacity of AI technologies to carry out several occupational tasks. In making your judgement for each task:

- Please consider “current” computer techniques [...].
- Please consider techniques that might need “reasonable advance preparation”. You can consider possible AI systems involving any level of development effort as long as the work involves established AI techniques.

Analyses of the results and comments obtained from experts in the exploratory studies highlighted a better understanding of the task and the type of AI systems they should consider than in previous studies. However, group discussions and feedback on ratings were still necessary to remove remaining misunderstandings and share additional precisions on the AI systems they envisaged.

Question phrasing and response format

The project also explored different possibilities of asking the questions and response formats and their implications for the reliability of experts’ judgements, and the analysis and interpretation of data. The expert meeting organised in March 2021 discussed the advantages and disadvantages of:

- simple categorical questions (Yes/No/Maybe)
- Likert scale questions with probabilities of whether AI systems can solve the question (with or without detailed rubrics for each level on the scale)

- open-ended questions to elicit the rationale of expert judgements.

In addition, the project considered ways to elicit experts' confidence in their judgement, which can be important information to communicate to the policy community. Experts at the March 2021 meeting (including computer scientists and psychologists with survey expertise) endorsed the use of a scale that simultaneously captured experts' judgement of AI capabilities and confidence in their judgements. The question "*How confident are you that your AI system could carry out this task?*" with a **Likert or continuous scale of probabilities** and a "Don't know" option received overall positive feedback from experts. The analysis of the quantitative results and how the subsequent meeting helped finetune experts' judgement and increase levels of certainty, are discussed in Chapter 3.

Experts also agreed to provide **rationales and comments** following their answers. This allowed them to express uncertainty and complement their answers with clarifications and/or references. Such qualitative information was also valuable for the team to better understand quantitative judgements and to prepare the group discussions following the ratings. Table 2.4 summarises the instructions and response formats used across the exploratory studies.

Table 2.4. Task framing and response format in the AIFS project

	Task framing and instructions	Response format (scale)
PIAAC 2016	<ul style="list-style-type: none"> • No framework document • USD 1M + 1 year development 	<ul style="list-style-type: none"> • Yes / No / Maybe • No rationale
PIAAC 2021 follow-up 1	<ul style="list-style-type: none"> • Framework document • USD 1M + 1 year development 	<ul style="list-style-type: none"> • Discrete probabilities (0%; 25%; 50%; 75%, 100%) • Rationale
PIAAC 2022 follow-up 2	<ul style="list-style-type: none"> • Framework document • Imagined AI system based on existing techniques 	<ul style="list-style-type: none"> • Discrete probabilities (0%; 25%; 50%; 75%, 100%) • Rationale
PISA 2022 core experts	<ul style="list-style-type: none"> • Framework document • Imagined AI system based on existing techniques 	<ul style="list-style-type: none"> • Discrete probabilities (0%; 25%; 50%; 75%, 100%) • Rationale
PISA 2022 new experts	<ul style="list-style-type: none"> • Framework document • Imagined AI system based on existing techniques 	<ul style="list-style-type: none"> • Continuous probabilities (0-100%) • Rationale
Occupational 2022	<ul style="list-style-type: none"> • Framework document • Current techniques with reasonable advanced preparation 	<ul style="list-style-type: none"> • Discrete probabilities (0%; 25%; 50%; 75%, 100%) • Rationale

Finally, to track technological advances and evolution over time, the 2016 pilot study asked experts about the projected capability of AI systems solving similar tasks over the short term (5 years) and long term (10 to 20 years). Experts felt more confident about short-term projections as they could link them to ongoing research projects. In addition, grant applications typically require five-year projections (Elliott, 2017^[11]). Projections provide comparative data (an assessment in five years can be compared to projections), particularly for longer periods. Questions about future AI capabilities in the exploratory studies were limited to a five-year projection for the PIAAC 2021 follow-up 2022 assessments.

Establishing consensus: Quantitative disagreement versus qualitative agreement

As one of its most important objectives, the AIFS exploratory studies tested whether and with what method it is possible to establish consensus among experts. Consensus can be an indicator of data quality and the usefulness of expert judgement to inform policy decisions.

Consensus or agreement among experts can be measured via quantitative and qualitative methods (for a full review, see von der Gracht (2012^[6])). The AIFS project primarily used simple mathematical aggregations and comparisons of ratings as quantitative methods:

- Simple and two-thirds majority: more than half (or two-thirds) of experts gave the same rating (e.g. said “Yes, AI can solve this task”). Can be adapted to discrete or continuous scales by setting a threshold-point for decision.
- Interquartile range (IQR), standard deviations, coefficients of variation: measures of dispersion. The higher the value is, the more ratings are spread around the mean or median. A low value can be an indicator of consensus across experts’ ratings.

These simple measures provided an effective way to compare experts’ ratings across the different assessment scales and allowed for a straightforward analysis and interpretation of results. Other, more complex measures can be used, such as Kappa and Kendall’s coefficient of concordance. Kappa in the exploratory studies generally indicated low levels of agreement.

The project revisited methods for collecting and aggregating expert judgement to increase consensus. This was only partially achieved from the 2016 pilot study to the 2021/22 assessments: agreement on the literacy questions increased but not on the numeracy questions. Importantly, there was still no overall consensus among experts: the ratings showed considerable variations (see Chapter 3 and Chapter 5) in any of the exploratory studies. This could be partly due to the difference between raters’ domain of expertise, which affects their judgements. For example, NLP experts might not be aware of all the technological advances in the vision domain. This, in turn, could negatively bias their judgements on questions involving computer vision. As another explanation for lack of consensus, information included in the task framing for experts was still not enough for a common understanding of how AI capabilities on the questions should be rated. Although additional information in task framing could help increase consensus, there are practical limits in the amount of preparatory information that respondents are willing to review when the rating task itself is already quite long.

Group discussions and an *analysis of experts’ rationales* have provided substantive qualitative data to understand consensus/dissensus among experts and identify their reasons. In the 2016 PIAAC group discussion, experts agreed on common challenges of current AI systems, such as the difficulty to deal with multimodal questions or the likely overfitting of systems (i.e. systems are fine-tuned to solve a specific set of questions) (Elliott, 2017^[1]). The 2021/22 follow-up assessments showed a stronger consensus on several aspects of AI state of the art that became apparent in the group discussions. In the PIAAC 2021 repeat, the quantitative analysis of expert ratings showed disagreement across experts on AI capabilities to solve the numeracy questions. However, the rationales provided in the survey and the group discussion showed overall agreement about AI systems’ capabilities to solve the PIAAC numeracy questions (OECD, 2023^[20]).

The discrepancy between the level of consensus in the quantitative and qualitative analysis of experts’ judgements can be largely explained by two factors. First, the limitations of task framing described above (see Chapter 3). Second, the differences in computer scientists’ domain expertise and knowledge of the latest AI performance measures. Experts tended to base their judgements on the direct measures of AI performance that they know and that are relevant to the given set of questions. Naturally, one expert cannot know all the thousands of such measures and cannot follow their rapid evolution. However, interactions during the follow-up meeting allowed them to exchange references and reconsider their judgement in view of the evidence shared by others.

Overall, the exploratory studies have shown that despite several revisions and improvements, reaching consensus was not possible based on a purely quantitative analysis of expert judgements. When quantitative analysis was complemented with qualitative information, however, a global consensus was possible in most cases.

Conclusions: Challenges and future directions

This chapter described the processes and methods for engaging experts and collecting their judgement on AI capabilities using human tests/tasks. The refinements to the methodology since the pilot study explored only a small set of configurations discussed in the literature. Nevertheless, the explorations highlighted some key challenges of this approach.

First, collecting expert judgements on such complex assessments proved to be more resource-intensive than anticipated. This was true both in terms of financial costs and the time commitment required from experts and the project team. Part of the problem is the natural limit on the number of people with both the appropriate expertise and the interest to engage with this work. Identifying enough experts and raising their interest to engage with this work require substantial time from the project team. The project tested various incentives to engage experts and found that some (particularly money and a combination of several incentives) work but are costly.

The project's goal is to regularly update the measures of AI capabilities once they are developed to inform the policy community. Using human tests/tasks, such as the OECD's educational tests (PIAAC, PISA) and occupational tasks, means the project would need to collect expert judgements regularly (e.g. every two-five years). The methodological explorations described above indicate this will be very difficult, if feasible at all with a large number of experts given the limited interval between assessments and the resources available. On the positive side, the EKE literature and the exploratory assessments suggest a smaller group of experts' judgements gives similar aggregate results to that of a larger group (see Chapter 3). However, the team has recognised that engaging even a smaller group of experts on a regular basis would be substantially more time-consuming and expensive than originally believed.

Second, it is challenging to formulate tasks that provide valid and reliable expert judgement and yield an acceptable level of quantitative consensus in cases where experts agreed qualitatively. The project worked with experts to reflect on the task framing and instructions and improved its methodology through multiple exploratory assessments. Despite trying several different techniques and achieving expert agreement on qualitative descriptions of current AI capabilities, the project could not get adequate agreement in experts' quantitative judgements of those capabilities.

Overall, the explorations concluded that using expert judgement to establish measures of AI capabilities has limits. The project therefore began to explore using direct measures of AI systems originating from benchmark tests, competitions and formal evaluations. This seemed to be a natural choice for two reasons. First, a huge amount of such measures exists in the field of computer science and they are constantly growing in number. Second, experts participating in the exploratory studies continuously referred to such direct measures when making their judgements. Thus, it was straightforward to rely directly on these measures instead of their judgements. Despite the shift of focus from expert judgement to synthesising direct measures, the former remains relevant in certain domains where direct measures are not available.

Alternative pathways to develop AI measures led the project to rely on experts in different ways. Experts' role in identifying and interpreting the results of available direct measures became stronger than providing their judgements about likely performance on specific tasks. New roles involve experts from different domains to work together, build a shared understanding of the project goal and collectively develop tools. Examples of such co-construction are the development of the facets to characterise benchmarks (see Chapter 6), the classification of formal evaluation campaigns (see Chapter 7), the design of the occupational assessment (Chapter 5) and the development of AI capability scales (Chapter 9).

Over the past three years, the AIFS project has developed a core group of committed experts and a set of methods that allow for obtaining valid and reliable expert judgements across several domains. The rest of the report will describe in more detail the exploratory assessments and other approaches to summarise the state-of-the-art AI capabilities.

References

- Anderson, B. (ed.) (2018), “A practical guide to structured expert elicitation using the IDEA protocol”, *Methods in Ecology and Evolution*, Vol. 9/1, pp. 169-180, <https://doi.org/10.1111/2041-210x.12857>. [15]
- Bolger, F. and G. Rowe (2015), “The aggregation of expert judgment: Do good things come to those who weight?”, *Risk Analysis*, Vol. 35/1, pp. 5-11, <https://doi.org/10.1111/risa.12272>. [12]
- Bolger, F. and G. Rowe (2015), “There is data, and then there is data: Only experimental evidence will determine the utility of differential weighting of expert judgment”, *Risk Analysis*, Vol. 35/1, pp. 21-26, <https://doi.org/10.1111/risa.12345>. [13]
- Cooke, R. (1991), *Experts in Uncertainty: Opinion and Subjective Probability in Science*, Oxford University Press. [3]
- Delbecq, A. and A. Van de Ven (1971), “A group process model for problem identification and program planning”, *The Journal of Applied Behavioral Science*, Vol. 7/4, pp. 466-492, <https://doi.org/10.1177/002188637100700404>. [16]
- Delbecq, A., A. Van de Ven and D. Gustafson (1975), *Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes*, Scott Foresman and Company. [17]
- 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]
- Freedman, J. and S. Fraser (1966), “Compliance without pressure: The foot-in-the-door technique”, *Journal of Personality and Social Psychology*, <https://doi.org/10.1037/h0023552>. [19]
- Gosling, J. (2018), “SHELF: The Sheffield Elicitation Framework”, in *International Series in Operations Research & Management Science, Elicitation*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-65052-4_4. [18]
- Hanea, A. et al. (eds.) (2021), *Expert Judgement in Risk and Decision Analysis*, Springer Cham, <https://doi.org/10.1007/978-3-030-46474-5>. [4]
- Kahneman, D., P. Slovic and A. Tversky (eds.) (1982), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press. [7]
- Lichtenstein, S., B. Fischhoff and L. Phillips (1982), “Calibration of probabilities: The state of the art to 1980”, in *Judgment under Uncertainty*, Cambridge University Press, <https://doi.org/10.1017/cbo9780511809477.023>. [9]
- Linstone, H. and M. Turoff (eds.) (1976), *The Delphi Method: Techniques and Applications*, Addison-Wesley, <https://doi.org/10.2307/3150755>. [8]
- Myers, D. and H. Lamm (1976), “The group polarization phenomenon.”, *Psychological Bulletin*, Vol. 83/4, pp. 602-627, <https://doi.org/10.1037/0033-2909.83.4.602>. [10]
- 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>. [20]

- O'Hagan, A. et al. (2006), *Uncertain Judgements: Eliciting Expert Probabilities*, John Wiley. [2]
- Rowe, G. (1992), "Perspectives on Expertise in the Aggregation of Judgments", in G. Wright and F. Bolger (eds.), *Expertise and Decision Support*, Plenum, https://doi.org/10.1007/978-0-585-34290-0_8. [5]
- Rowe, G., G. Wright and F. Bolger (1991), "Delphi: A reevaluation of research and theory", *Technological Forecasting and Social Change*, Vol. 39/3, pp. 235-251, [https://doi.org/10.1016/0040-1625\(91\)90039-j](https://doi.org/10.1016/0040-1625(91)90039-j). [14]
- Turner, M. and A. Pratkanis (1998), "Twenty-five years of groupthink theory and research: Lessons from the evaluation of a theory", *Organizational Behavior and Human Decision Processes*, Vol. 73/2-3, pp. 105-115, <https://doi.org/10.1006/obhd.1998.2756>. [11]
- von der Gracht, H. (2012), "Consensus measurement in Delphi studies", *Technological Forecasting and Social Change*, Vol. 79/8, pp. 1525-1536, <https://doi.org/10.1016/j.techfore.2012.04.013>. [6]

Notes

¹ To ensure ethical treatment, after completing the survey, all respondents received both the EUR 800 honorarium and co-authorship.

3 Assessing AI capabilities with education tests

Mila Staneva (OECD), Abel Baret (OECD), Àngel Aso-Mollar (Universitat Politècnica de València), Joseph Blass (Northwestern University), Salvador Carrión Ponz (Universitat Politècnica de València), Vincent Conitzer (Carnegie Mellon University), Ulises Cortes (Universitat Politècnica de Catalunya), Pradeep Dasigi (Allen Institute for AI), Angel de Paula (Universitat Politècnica de València), Carlos Galindo (Universitat Politècnica de València), Janice Gobert (Rutgers University), Jordi González (Universitat Autònoma de Barcelona), Fredrik Heintz (Linköping University), Jim Hendler (Rensselaer Polytechnic Institute), Daniel Hendrycks (Center for AI Safety), Lawrence Hunter (University of Colorado Anschutz Medical Campus), Juan Izquierdo-Domenech (Universitat Politècnica de València), Maria Juarez (Universitat Politècnica de València), Aina Juraco Frias (Universitat Politècnica de València), Aviv Keren (Anyword); Rik Koncel-Kedziorski (Kensho Technologies), David Leake (Indiana University), Bao Sheng (Aiden) Loe (University of Cambridge), Fernando Martinez-Plumed (Universitat Politècnica de València), Aqueasha Martin-Hammond (Indiana University), Cynthia Matuszek (University of Maryland, Baltimore County), Antoni Mestre Gascón (Universitat Politècnica de València), Jose Andres Moreno (Universitat Politècnica de València), Constantine Nakos (Northwestern University), Taylor Olson (Northwestern University), Carolyn Rose (Carnegie Mellon University), Areg Mikael Sarvazyan (Universitat Politècnica de València), Brian Scassellati (Yale University), Wout Schellaert (Universitat Politècnica de València), Claes Strannegård (Chalmers University of Technology), Neset Tan (University of Auckland), Tadahiro Taniguchi (Ritsumeikan University), Karina Vold (University of Toronto), Michael Wooldridge (University of Oxford)

This chapter introduces three exploratory studies that assessed the capabilities of artificial intelligence (AI) through standardised education tests designed for humans. The first two studies, conducted in 2016 and 2021/22, asked experts to evaluate AI's performance on the literacy and numeracy tests of the OECD's Survey of Adult Skills (PIAAC). The third study collected expert judgements of whether AI can solve science questions from the OECD's Programme for International Student Assessment (PISA). The studies aimed to refine the assessment framework for eliciting expert knowledge on AI using established educational assessments. They explored different test formats, response methodologies and rating instructions, along with two distinct assessment approaches. A "behavioural approach" used in the PIAAC studies emphasised smaller expert groups engaging in discussions, and a "mathematical approach" adopted in the PISA study relied more heavily on quantitative data from a larger expert pool. This chapter presents the results of the studies and discusses the advantages and disadvantages of their methodological approaches.

The AI and the Future of Skills (AIFS) project carried out three exploratory studies on the use of education tests for collecting expert evaluations on artificial intelligence (AI). The first two studies used the OECD's Survey of Adult Skills of the Programme for International Assessment of Adult Competencies (PIAAC). PIAAC assesses the proficiency of adults aged 16-65 in three general cognitive skills – literacy, numeracy and problem solving in technology-rich environments (OECD, 2019_[1]).¹ In 2016, a pilot study asked 11 experts to assess whether AI can do the literacy, numeracy and problem-solving tests of PIAAC (Elliott, 2017_[2]). This pilot study served as a stepping stone into the AIFS project. In 2021/22, a follow-up study surveyed 15 computer experts to show how AI capabilities in literacy and numeracy have evolved since the pilot assessment (OECD, 2023_[3]).

A third study in 2022 used test questions from the OECD's Programme for International Student Assessment (PISA) as a measurement tool. PISA assesses the knowledge and skills of 15-year-old students in reading, mathematics and science (OECD, 2019_[4]). The study collected expert judgement on AI capabilities using questions from the science domain. In contrast to the studies using PIAAC, this study attempted to assemble a much larger sample of experts. For this purpose, the study offered different incentives for attracting and engaging experts in the assessment (see Chapter 2).

The three exploratory studies aimed to improve the assessment framework for eliciting expert knowledge on AI using standardised tests designed for humans. To that end, the studies explored the use of different tests, different response formats and different instructions for rating. Moreover, they tested the feasibility of two generally different approaches to assessing expert knowledge for the project purposes. The studies using PIAAC explored the so-called behavioural approach (see Chapter 2). This means they relied on smaller expert groups that could engage in extensive discussions to reach agreement on AI capabilities. By contrast, the study using PISA followed a so-called mathematical approach. This means it relied more heavily on quantitative information from a larger group of experts, under the assumption that aggregation across many judgements reduces random error in those judgements.

The use of education tests for assessing AI capabilities can provide both reliable and policy-relevant AI measures. Education tests provide standardised and objective criteria for assessing AI capabilities. This enables assessing AI with different expert groups and tracking AI progress across time. Moreover, education tests typically target skills that are taught in education institutions and widely used at work. Assessing how AI performs on these skills thus provides insights into AI's potential impacts on education and employment. This information is important for designing education and labour market policies that are responsive to incoming technological changes.

The results of the exploratory studies showed that AI performs well in all three tested domains. In literacy, computer experts expected that AI could solve 80% of the PIAAC questions in 2021. This marks a considerable improvement to the success rate of 55% in literacy obtained in 2016. In numeracy, AI was expected to solve around two-thirds of the PIAAC test questions in 2021/22. In science, experts expected AI to solve PISA science questions with 61% confidence. These results correspond to human performance levels in the middle or at the higher end of the performance spectrum. They show that AI can potentially outperform large shares of the adult and youth population. AI's estimated performance in literacy, for example, is equal to or higher than that of 90% of adults in OECD countries with PIAAC data (OECD, 2023_[3]).

However, the results also revealed some methodological challenges in collecting expert judgements on AI capabilities with education tests. The main challenge was disagreement among experts, especially in rating AI's potential performance on the PIAAC numeracy test. Disagreement mainly related to ambiguity about how general the computer capabilities being assessed are supposed to be. Some experts assumed general capabilities that should enable successful performance over a wide range of test questions. Others considered narrow systems geared towards solving specific problems. To reach agreement, the experts thus needed clarification on the generality of the underlying capabilities being evaluated. The studies explored different methods to address this issue.

This chapter describes the three exploratory studies and compares their methodologies. The first section discusses the advantages of using education tests to collect expert judgements on AI. The second section introduces the survey instruments. The third section presents the methodology – the methods used for expert knowledge elicitation, including the questionnaires used for the expert surveys. The fourth section presents the main findings and discusses the quality of measures. The fifth section elaborates on the strengths and weaknesses of the approaches used and the sixth section outlines the next steps in this project strand.

Rationale for assessing AI capabilities with education tests

The elicitation of expert ratings on AI capabilities with education tests has a number of methodological advantages:

- Education tests provide a standardised and objective way for eliciting expert knowledge. This enables reproducing an AI assessment with different groups of experts and across time.
- Education tests provide concrete and detailed descriptions of the test tasks. This enables computer experts to make more objective and precise judgements since they do not have to rely on implicit assumptions about the task requirements. This improves the reliability of the assessment.
- Education tests enable comparisons of computer and human capabilities. This can show which skills may become obsolete and which may gain in significance in the years ahead. Moreover, education tests typically assess various socio-demographic characteristics of respondents in addition to their skill proficiency. This enables fine-grained AI-human comparisons within different country contexts, age groups or occupations. Such analysis can show which social groups are particularly vulnerable to automation.
- Education tests offer a graduated progression from simple to complex tasks. This allows for obtaining more nuanced measures of AI performance across different levels of task difficulty.
- Assessing AI capabilities on education tests provides information that is useful for policy making. Both PIAAC and PISA assess key cognitive skills that are used in most social contexts and work situations. These skills strongly affect individuals' ability to participate effectively in the labour market, education and training, and social and civic life (OECD, 2019^[1]). Understanding how AI performs with respect to these skills can thus provide valuable insights into AI's potential impacts on work and life.
- Assessing AI against education tests provides understandable measures. Compared to benchmark tests and evaluations used in AI research, education tests can describe AI capabilities in a way that is meaningful to the general public. In addition, educators and education researchers are typically familiar with the skills assessed in education tests and the ways those skills are developed in education and used at work and in daily life.

Against this background, expert judgement on whether AI can carry out education tests constitutes an important source of information for the AIFS project. This information can complement the overall assessment framework in areas in which results from direct assessments of AI systems are lacking.

Box 3.1. Use of education tests in AI evaluation

Computer scientists employ various education tests to directly assess AI systems' performance. For instance, Hendrycks et al. (2020^[5]) evaluated state-of-the-art AI models, including different configurations of GPT-3 (Generative Pre-trained Transformer), on 57 education tests. The tests cover various disciplines, including mathematics, physics, history, micro econometrics, geography, law, anatomy and philosophy. They span elementary to university-level courses. While most models performed at nearly random levels, the largest GPT-3 achieved an average accuracy of 44% across tests. This performance was lowest in subjects that require quantitative reasoning, such as mathematics and physics, and in subjects related to human values, such as law and ethics.

AI performance on education tests has increased with the introduction of more powerful language models. GPT-3.5, introduced in early 2022, demonstrated strong performance in college-level art history, environmental science, psychology, political studies and writing, among others (OpenAI, 2023^[6]). GPT-4, introduced in March 2023, outperformed GPT-3.5 on most of these exams (Bubeck et al., 2023^[7]; OpenAI, 2023^[6]). However, performance in quantitative subjects remains moderate. A study that evaluated the mathematical capabilities of GPT-4 concluded that while the model performs well in undergraduate-level mathematics, it often fails on graduate-level difficulty (Frieder et al., 2023^[8]).

Some computer scientists have argued that standardised education tests are a useful evaluation tool for AI systems. According to Clark and Etzioni (2016, p. 4^[9]), such tests are “accessible, easily comprehensible, clearly measurable, and offer a graduated progression from simple tasks to those requiring deep understanding of the world”. Additionally, they encompass a broad spectrum of AI capabilities. Hendrycks et al. (2020^[5]) note that education tests require extensive world knowledge and problem solving ability. They thus provide important insights into AI models' overall language understanding abilities.

Overview of the education tests used

The following subsections introduce PIAAC and PISA. They provide information on the approaches these surveys use to assess skills and describe the formats of test questions, as well as the contexts and cognitive strategies they address.

Assessing literacy and numeracy in the Survey of Adult Skills (PIAAC)

The Survey of Adult Skills (PIAAC) is conducted every ten years. The First Cycle took place between 2011 and 2018, collecting data from 39 countries and economies. It surveyed approximately 250 000 respondents, with national samples ranging from about 4 000 to nearly 27 300 (OECD, 2019^[1]). First results from the Second Cycle are expected in 2024.

The survey assesses the proficiency of adults aged 16-65 in literacy, numeracy and problem solving with computers. The pilot study from 2016 focused on all three domains, while the follow-up study from 2021/22 used only the literacy and numeracy assessments of PIAAC. This report covers only results on AI capabilities in literacy and numeracy since they were assessed in both time points. These skills are considered key cognitive skills since they build the foundation for developing higher-order skills (e.g. analytic reasoning and learning-to-learn skills) and for acquiring new knowledge. In technology-rich societies, literacy and numeracy are essential for gaining access to information relevant to everyday life (OECD, 2012^[10]).

PIAAC assesses both skill proficiency and question difficulty on a 500-point scale. Questions are first assigned a difficulty score, which is dependent on the proportion of respondents who complete them successfully. Respondents are then placed on the same scale, based on the number and difficulty of questions they answer correctly. For simplicity, the 500-point scale is broken down into six proficiency/difficulty levels (below Level 1, Levels 1-5). A person with a proficiency score in the middle of the range defining the level can successfully complete tasks at that level approximately 67% of the time. For example, a person with a score in the middle of Level 2 would score close to 67% in a test made up of items of Level 2 difficulty (OECD, 2013_[11]).

The PIAAC literacy test measures adults' ability to understand, evaluate, use and engage with written texts in real-life situations. It contains 57 reading tasks that adults typically encounter in work and personal life. Examples include job postings, webpages, newspaper articles and e-mails. These texts are presented in different formats – as print texts, digital texts, continuous texts, sentences formed into paragraphs or non-continuous texts, such as those appearing in charts, lists or maps. Items can also contain multiple texts that are independent from each other but linked for a particular purpose (OECD, 2012_[10]; OECD, 2013_[11]).

Easy literacy tasks (below Level 1 and Level 1) require knowledge and skills in recognising basic vocabulary and reading short texts. Tasks typically require the respondent to locate a single piece of information within a brief text. In intermediate-level tasks (Levels 2 and 3), understanding text and rhetorical structures becomes more central, especially navigating complex digital texts. Texts are often dense or lengthy. They may require the respondent to construct meaning across larger chunks of text or perform multi-step operations to identify and formulate responses. Hard tasks (Levels 4 and 5) require complex inferences and application of background knowledge. Texts are complex and lengthy and often contain competing information that is seemingly as prominent as correct information. Many tasks require interpreting subtle evidence-based claims or persuasive discourse relationships.

The PIAAC numeracy test measures the ability to access, use, interpret and communicate mathematical information and ideas to manage the mathematical demands of everyday life (OECD, 2012_[10]; OECD, 2013_[11]). It contains 56 tasks that are designed to resemble real situations from work and personal life, such as managing budgets and project resources, and interpreting quantitative information presented in the media. The mathematical information can be presented in many ways, including images, symbolic notations, formulae, diagrams, graphs, tables and maps. Mathematical information can be further expressed in textual form (e.g. “the crime rate increased by half”).

Easy numeracy tasks (below Level 1 and Level 1) require respondents to carry out simple, one-step processes. Examples are counting, understanding simple percentages or recognising common graphical representations. The mathematical content is easy to locate. Tasks at medium difficulty levels (Levels 2 and 3) require the application of two or more steps or processes. This can involve calculation with decimal numbers, percentages and fractions, or the interpretation and basic analysis of data and statistics in texts, tables and graphs. The mathematical information is less explicit and can include distractors. Hard tasks (Levels 4 and 5) require understanding and integrating multiple types of mathematical information, such as statistics and chance, spatial relationships and change. The mathematical information is presented in complex and abstract ways or is embedded in longer texts.

Box 3.2. Example items from PIAAC and PISA

Figure 3.1 presents example items from the PIAAC literacy and numeracy tests. These sample items are at difficulty Level 3.

Figure 3.1. PIAAC Literacy and Numeracy – Sample items

Panel A literacy

Unit 1 - Question 1/3

Look at the list of preschool rules. Highlight information in the list to answer the question below.

What is the latest time that children should arrive at preschool?

Preschool Rules

Welcome to our Preschool! We are looking forward to a great year of fun, learning and getting to know each other. Please take a moment to review our preschool rules.

- Please have your child here by 9:00 am.
- Bring a small blanket or pillow and/or a small soft toy for naptime.
- Dress your child comfortably and bring a change of clothing.
- Please no jewelry or candy. If your child has a birthday please talk to your child's teacher about a special snack for the children.
- Please bring your child fully dressed, no pajamas.
- Please sign in with your full signature. This is a licensing regulation. Thank you.
- Breakfast will be served until 7:30 am.
- Medications have to be in original, labeled containers and must be signed into the medication sheet located in each classroom.
- If you have any questions, please talk to your classroom teacher or to Ms. Marlene or Ms. Tree.

Panel B numeracy

Look at the graph about the number of births. Click to answer the question below.

During which period(s) was there a decline in the number of births? Click all that apply.

1957 - 1967

1967 - 1977

1977 - 1987

1987 - 1997

1997 - 2007

The following graph shows the number of births in the United States from 1957 to 2007. Data are presented every 10 years.

Year	Number of Births
1957	4,300,000
1967	3,520,959
1977	3,326,632
1987	3,809,394
1997	3,880,894
2007	4,315,000

Source: OECD (2012^[10]), *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*, <http://dx.doi.org/10.1787/9789264128859-en>.

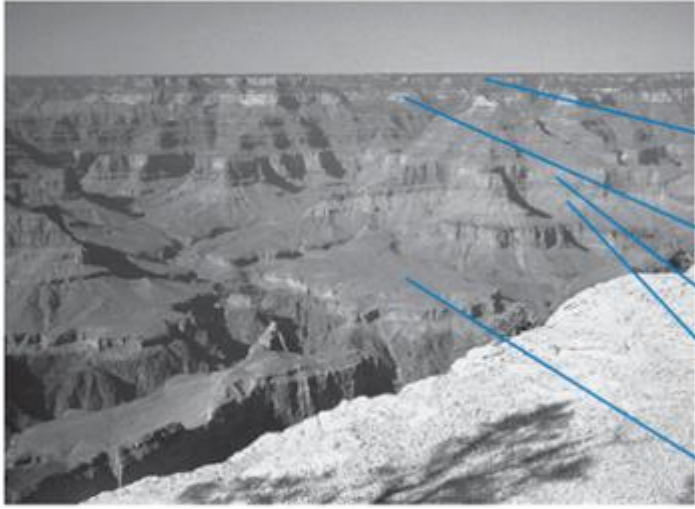
Figure 3.2 shows an example for an item of the PISA science test at difficulty Level 2.

Figure 3.2. PISA Science – Sample item

SCIENCE UNIT 7: THE GRAND CANYON

The Grand Canyon is located in a desert in the USA. It is a very large and deep canyon containing many layers of rock. Sometime in the past, movements in the Earth's crust lifted these layers up. The Grand Canyon is now 1.6 km deep in parts. The Colorado River runs through the bottom of the canyon.

See the picture below of the Grand Canyon taken from its south rim. Several different layers of rock can be seen in the walls of the canyon.



Limestone A

Shale A

Limestone B

Shale B

Schists and granite

QUESTION 7.1

The temperature in the Grand Canyon ranges from below 0 °C to over 40 °C. Although it is a desert area, cracks in the rocks sometimes contain water. How do these temperature changes and the water in rock cracks help to speed up the breakdown of rocks?

- A. Freezing water dissolves warm rocks.
- B. Water cements rocks together.
- C. Ice smooths the surface of rocks.
- D. Freezing water expands in the rock cracks.

Source: OECD (2009_[12]): *Take the Test: Sample Questions from OECD's PISA Assessments*, <https://doi.org/10.1787/9789264050815-en>.

Assessing science literacy in the Programme for International Student Assessment (PISA)

PISA assesses the knowledge and skills of 15-year-old students in reading, mathematics and science. The assessment has taken place every three years since 2000, with each round testing one of the three subjects in detail and providing basic results for the other two. In addition, some assessment rounds offer optional assessments, for example, on students' familiarity with information and communication technologies (ICT) in 2015 (OECD, 2016_[13]) or on students' well-being in 2018 (OECD, 2019_[4]). The last assessment round to date, in 2022, collected information from more than 80 countries and economies.

The AIFS project focused on the PISA science assessment. It used 20 publicly released test questions from the science domain, which were either used in actual assessments or tested in PISA field trials (see Figure 3.2 in Box 3.2 for an example).² PISA's science questions measure students' scientific knowledge, as well as the ability to use that knowledge to identify scientific issues, explain phenomena scientifically or use scientific evidence. They use multiple-choice or open-ended response formats and are typically presented in text, although some questions contain images, graphics or tables. The questions are designed to resemble a wide variety of real-life situations that involve science and technology. Topics are related to personal (e.g. nutrition), social (e.g. disease control) or global (e.g. climate change) issues and cover the domains of health, natural resources, environmental quality, hazards and the frontiers of science and technology.

Scores in PISA are scaled to fit a normal distribution with a mean of 500 score points and a standard deviation of 100 score points. They do not have a substantial meaning and, theoretically, there is no minimum or maximum score (OECD, 2019^[4]). In the first detailed assessment of science in 2006, around two-thirds of students in OECD countries scored between 400 and 600 score points (OECD, 2007^[14]). Similarly, as in PIAAC, the difficulty of individual questions is given a score on the scale that depends on the proportion of test takers getting the question correct. Student performance is then described by assigning each student a score according to the number and difficulty of questions he or she has answered correctly.

To ease the interpretation of results, PISA summarises science scores into six proficiency levels (Levels 1 to 6). Questions at Levels 1 and 2 are easier, requiring students to recall simple scientific facts or to use common scientific knowledge in drawing or evaluating conclusions. Questions at medium difficulty at Levels 3 and 4 require students to use scientific knowledge to make predictions, provide explanations, recognise relevant questions, and select relevant information from competing data or claims. Hard tasks at Levels 5 and 6 require students to create or use conceptual models to predict or explain scientific phenomena; to understand the design of scientific studies and the hypotheses they test; to use data to evaluate alternative viewpoints or differing perspectives; and to communicate scientific results.

Methodology for collecting expert judgement on AI with education tests

The pilot study using PIAAC was carried out with 11 experts in May 2016 (Elliott, 2017^[21]). Five years later, a second study followed up, using a comparable revised methodology (OECD, 2023^[33]). This follow-up study consisted of two rounds. In December 2021, 11 computer experts completed the literacy and numeracy assessments. Due to diverging ratings in the numeracy domain, a second round of interviews in September 2022 engaged four computer scientists with expertise in mathematical reasoning of AI. This second round applied a modified framework for rating to address expert disagreement.

In 2022, a third study using PISA questions was carried out. The study used the latest modified framework for rating. It first collected judgements from 12 core experts who participated in the previous PIAAC assessment. The study then conducted a large-scale online assessment with new experts to compare the advantages of this approach to the use of smaller groups of experts who engage with the project on a long-term basis. The following sections describe the methodology used in the three studies and how it was improved in the course of the work.

Collecting expert judgement

The methodology for collecting expert judgement in the three exploratory studies progressed from a behavioural to a mathematical approach (Rowe, 1992^[15]). As described in Chapter 2, the behavioural approach relies on a few experts who engage in in-depth discussions to arrive at a consensus judgement on a question. This aims to address questions in their complexity by considering different arguments and

perspectives and to draw on the best of these arguments to build a group judgement. By contrast, the mathematical approach relies on many experts who provide individual judgements without interacting with each other. The goal is to avoid social biases such as social conformity or dominance of influential individuals.

The pilot study with PIAAC came closest to a behavioural approach. Here, the 11 experts made their ratings during a two-day meeting. Materials, containing instructions and the PIAAC questions in literacy, numeracy and problem solving, were provided in advance. Experts were encouraged to study the questions and provide initial comments and reactions prior to the meeting. During the meeting, the experts provided their judgements and discussed salient questions, problematic issues or any ideas and arguments that group members brought up (Elliott, 2017^[2]).

The follow-up study followed a more structured approach. Experts received the PIAAC questions one week in advance. They had then two weeks to provide their ratings in an online survey. During this time, they were able to access the survey at any time via an individualised survey link. Finally, the experts discussed the results in a subsequent four-hour online meeting (OECD, 2023^[3]).

Interaction between experts is a key element of these studies. In the pilot study, experts could freely discuss their evaluations and any other matters related to the assessment. In the follow-up study, experts could communicate with the group via e-mail at any point of the assessment process. Most importantly, they received feedback on the group results from the online survey. During the four-hour workshop, they discussed these results, focusing on questions that received diverging ratings. Afterwards, experts had the opportunity to revise their ratings in response to the feedback received and the group discussion. This interaction was intended to encourage information sharing. Since some experts may have more information on specific AI applications or recent research results, they should be able to share their knowledge with the group.

The study using PISA tested an approach where experts completed the online survey without the possibility to interact and without receiving the test materials in advance. The study started with replicating the approach used in the previous assessment with PIAAC. That is, 12 of the core experts were invited to participate in the study. They received the PISA science questions in advance, rated potential AI performance on them in an online survey and discussed the results in an online meeting. Subsequently, the study invited more than 180 new experts to participate in the AI assessment with PISA. Of these, 63 expressed interest in participating and 33 actually participated in the online survey. These new experts had one month to complete the online survey. During this time, they did not have contact with other participants.

This latter approach served three purposes. First, restricting interaction among group members should account for social biases. Such biases can occur, for example, when only ideas that are broadly acceptable to all group members are discussed, or when a charismatic person imposes his or her opinions on the group (Tversky and Kahneman, 1974^[16]). Second, surveying many experts should better represent opinions and expertise in the scientific community regarding AI capabilities. Third, the approach should offer a faster and less costly way of collecting expert judgement since experts are only completing the online survey.

Response categories

In the pilot study with PIAAC, experts rated whether AI could solve each of the test questions with a Yes, Maybe or No. The subsequent discussion revealed that experts differed in their interpretation of the Maybe category. Some experts used it to express genuine uncertainty about AI's performance, while others used it as a not very certain Yes (Elliott, 2017^[2]).

The follow-up study attempted to gather more nuanced information on the certainty of experts' answers. It used a different question to elicit expert knowledge: "How confident are you that AI technology can carry

out this task?”. The response options were “0% – No, AI cannot do it”, “25%”, “50% – Maybe”, “75%”, “100% – Yes, AI can do it” and “Don’t know”. This scale reflects both experts’ confidence and their rating of the capability of AI. For example, “0% No, AI cannot do it” means that experts are certain that AI cannot carry out the task, while 25% means that experts think that AI probably cannot do it. The “50% – Maybe” category means full uncertainty (OECD, 2023^[3]). The study using PISA assessed experts’ confidence in AI solving the task with the same question. However, the large-scale sample used a continuous scale ranging from 0% to 100% confidence.

Assessing uncertainty in experts’ answers is important for establishing more valid AI measures. Some experts may lack specific knowledge regarding AI’s capabilities on particular tasks. Others may have trouble understanding the test question or the instructions for rating. Accounting for this, for example, by giving uncertain ratings a lower weight in the analysis, can improve measures. Moreover, a high proportion of uncertain ratings on specific questions can draw attention to a lack of clarity of some tasks or to general ambiguity in the field regarding AI’s performance on the tasks. Indicating and excluding such problematic questions can improve the analysis.

Instructions for rating

In making their evaluations, experts needed to consider a hypothetical process of adapting current AI techniques to the specific context of the test questions as no AI systems are tailored for solving PISA or PIAAC. Therefore, existing systems should be adapted for these tests, for example, by training them on relevant examples or by coding information about specific vocabularies, relationships or types of knowledge representation, such as charts and tables. Experts should use identical parameters for this hypothetical development effort in order to provide consistent ratings.

The pilot study using PIAAC defined two such parameters for experts to consider. First, experts were instructed to think of “current” computer techniques, meaning any available techniques addressed sufficiently in the literature. That is, experts were asked to imagine applying available systems instead of creating entirely new ones. Second, the instructions asked experts to consider a development effort that costs up to USD 1 million and takes no longer than one year to implement (Elliott, 2017^[2]).

The follow-up study used the same criteria to define the boundaries of the hypothetical advance preparation of AI systems for the tests. However, after the first assessment round, experts suggested that these parameters should be revised. They generally saw the hypothetical investment of USD 1 million as insufficient and proposed fitting this effort to the size of a major commercial AI development project to better reflect reality in the field. In addition, experts pointed out that PIAAC questions have many and different response types, some of which may be difficult for computers (e.g. clicking an answer). They advised changing the instructions for rating to allow for some hypothetical transformation of the task format. Such transformation should remove trivial hurdles to solving the task with AI, without changing the nature of the capabilities the test attempts to measure (OECD, 2023^[3]). These suggestions were implemented both in the second round of the follow-up study with PIAAC and the study with PISA.

Framing the rating exercise

The studies using PIAAC instructed experts to imagine a single hypothetical AI system for solving each test domain. However, experts did not always follow this rule. Some viewed different question types within a test (e.g. numeracy questions containing tables) as separate, narrow problems and evaluated AI’s capacity to solve them independently from each other. That is, they considered different systems for different problems. By contrast, other experts viewed a test as a general challenge for AI to process multimodal inputs in various settings. They considered one system solving all test questions, including similar tasks that are not part of the test (OECD, 2023^[3]).

How experts saw the scope of the test affected their judgements. The ones who focused on narrow problems generally gave more positive ratings than those who focused on general challenges. This led to diverging evaluations. The divergence in experts' rating was most pronounced in the follow-up PIAAC numeracy assessment. The numeracy test contains more diverse question types, including graphs, images, tables and maps, compared to the literacy and science tests that consist mostly of text inputs. This increased the ambiguity about the types of question formats that a hypothetical system is supposed to master.

To address this issue, the studies needed to define the full range of problems that AI is supposed to solve in a test domain. However, providing such information is not trivial. It requires defining all types of tasks that humans who master the test are expected to solve, and that machines should also be able to solve to be assigned the same underlying capability. Therefore, several other steps were taken to improve expert agreement in the follow-up study with PIAAC and the study with PISA.

First, the studies provided experts with information from the assessment frameworks of PIAAC and PISA. These documents describe the conceptual frame of the assessments. They define the underlying skills targeted by the assessments and describe the types and formats of the test questions. This information was synthesised and supplemented by nine example survey questions of low, medium and high difficulty to exemplify the scope of the test and how general the capabilities required for solving it should be. It was provided to experts prior to the online survey in the assessment with the four experts in mathematical reasoning, and at the onset of the online survey in the large-scale assessment.

Second, the studies asked experts to describe a high-level approach for solving each test using the information from the assessment framework and the example tasks. Subsequently, they were asked to rate the potential success of their imagined approach on each question of the test. Encouraging experts to think of a single system that can tackle all problems in a test was intended to provide a common ground for the evaluation. It should also facilitate understanding and communication among experts since it enables them to review the arguments and considerations of their peers.

Additional questions

In addition to how experts assess AI on the test, the survey asked a number of other questions. All online surveys contained open-ended questions asking experts to explain their ratings of AI performance on each question. The goal was to collect additional qualitative information on the rationales behind the ratings. At the end of each survey, experts could report any difficulties in understanding or answering the questions in a domain or leave any comments or suggestions.

The studies also asked experts to predict the performance of AI on the tests in future. These projections were collected in order to explore ways of tracking AI progress over time. The pilot study using PIAAC asked experts to predict AI performance ten years in the future. The follow-up study using PIAAC and the study using PISA used a shorter time frame of five years. The discussion showed that experts are more confident in making predictions over a shorter time horizon given the rapid rate of change in AI technology. They are also used to projecting AI research trends over three to five years when applying for research grants.

One challenge in using tests developed for humans is that they take for granted capabilities that most humans share, such as short-term memory or object recognition. This may result in misleading AI measures if computers fail the tests because they lack such capabilities rather than because they lack the primary capabilities being assessed. To tackle this problem, the follow-up study using PIAAC included an additional question: "If you think that AI cannot carry out the entire task or you are uncertain about it, would you say that AI can carry out parts of the task? If so, which part(s)?" (OECD, 2023_[3]). This question was intended to specify the elements of tasks that are easy for machines to perform in order to collect more

precise information on computer performance. However, only a few experts made use of this question, which led the OECD team to abandon it in subsequent assessments.

Constructing aggregate measures

The studies used two aggregation approaches to construct single measures for AI literacy, numeracy and science performance from the individual expert ratings (OECD, 2023^[3]). The first approach relies on the majority opinion of experts. It labels each test question as solvable or not solvable by AI based on what most experts judged. Questions on which experts cannot reach majority agreement are excluded from the analysis. The aggregate measures of AI performance show the percentage share of test questions in a domain that AI could solve according to the majority of computer experts. These measures are comparable to human scores that show the expected probability of respondents of successfully completing test items. As another advantage, they are robust, relying on experts' consensus understanding of AI capabilities.

A second approach constructs final measures by averaging across all experts' ratings. That is, the aggregate AI measures are computed by taking the mean of experts' ratings on each question and then averaging these mean ratings across all questions in a domain. The advantage of these measures is that they reflect all experts' opinions about AI capabilities. However, they are harder to interpret and not comparable to human scores as they show the average confidence of experts that AI can solve the test.

The follow-up study with PIAAC used the "majority" rule to aggregate experts' ratings. This is in line with the behavioural approach for eliciting expert judgements that focuses on discourse and consensus building among experts. By contrast, the PISA science assessment, which follows the mathematical approach for expert knowledge elicitation, averages all experts' ratings to arrive at final AI measures. This reflects the goal of the mathematical approach to build measures representing a broad spectrum of expertise and opinions in the expert community.

In the following, results from the studies with PIAAC are presented by following the "majority" rule, while results from the PISA assessment are computed with the "average" rule. Annex 3.A presents analyses from each study using the alternative approach. All measures are presented for different levels of question difficulty to provide a more detailed picture of AI performance on the tests.

Results

This section outlines the assessed performance of AI on the PIAAC literacy, PIAAC numeracy and PISA science questions. It also evaluates the quality of these AI performance metrics through several indicators of validity and reliability.

To facilitate a more direct comparison, answer categories from the 2021/22 PIAAC literacy and numeracy assessments were aligned with the Yes, Maybe and No categories from the 2016 study. That is, ratings of 0% and 25% were combined into a No-category, and ratings of 75% and 100% were treated as Yes. The aggregate measures then show the share of test questions, for which the majority of experts give a Yes. In contrast, the AI measures obtained with PISA indicate the average level of experts' confidence in AI's ability to successfully complete the test tasks.

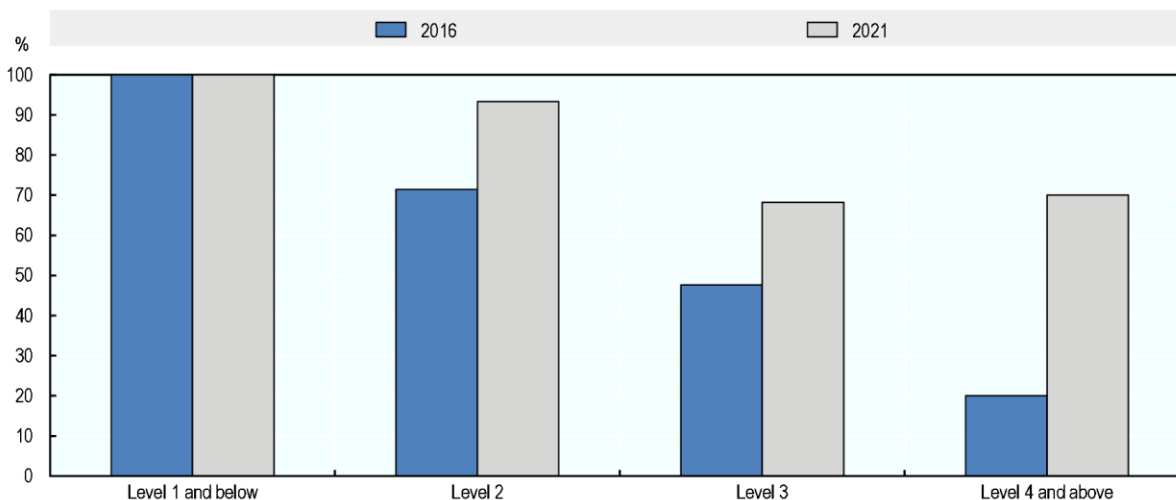
AI capabilities in literacy, numeracy and science

Figure 3.3 shows results from the pilot and follow-up studies with PIAAC for literacy. It indicates a clear improvement of AI literacy capabilities from 2016 to 2021. According to the majority of experts, AI's potential performance on the test increased at all difficulty levels. The increase amounts to 25 percentage points across all questions, moving from 55% to 80% between 2016 and 2021. These findings align well with the significant advances in natural language processing (NLP) that have occurred since 2016. These

include the advent of large pre-trained language models like GPT-2 and GPT-3, predecessors to ChatGPT (Radford et al., 2018_[17]). The coherence between experts' judgements and known progress in AI capabilities suggests that experts have a solid grasp of the task at hand.

Figure 3.3. AI literacy performance in 2016 and 2021, by question difficulty

Percentage share of literacy questions that AI can answer correctly according to a simple majority of experts; measures use Yes/No-ratings, Maybe omitted



Source: Adapted from (OECD, 2023_[3]), *Is Education Losing the Race with Technology?*, Figure 5.2, <https://doi.org/10.1787/73105f99-en>.

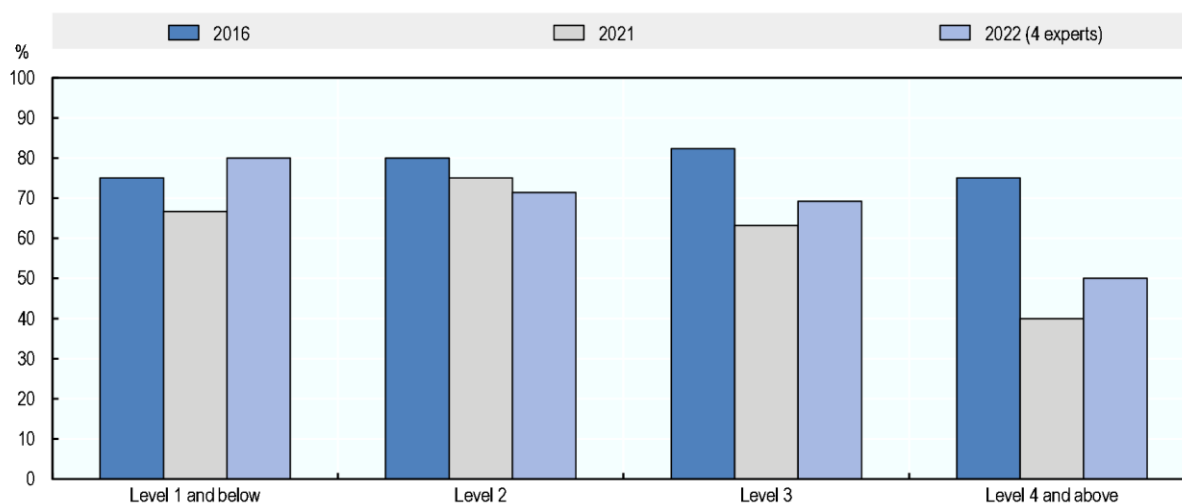
The results of the numeracy assessment are less straightforward. Following the same aggregation approach, Figure 3.4 shows a decline in AI's performance on the numeracy questions between 2016 and 2021/22. The decline is most pronounced at question difficulty Level 3 and Level 4 and above – 16 and 35 percentage points, respectively.

In the follow-up assessment, the 11 experts who completed the first assessment round in 2021 and the 4 mathematical reasoning experts who re-assessed numeracy in 2022 provided similar aggregate ratings. This suggests that neither the assessment modifications nor the shift in expertise significantly impacted group ratings on numerical skills.

These counter-intuitive results have to do with strong disagreement among experts in the follow-up study. Two opposing groups emerged. In the first round, five experts evaluated AI negatively on almost all questions, while four other experts provided mostly positive ratings. In the second round, one expert had overly negative ratings, another had mostly negative ratings and the other two were in the middle. This led to thin majorities, often determined by a single vote, and resulting in arbitrary conclusions on AI's capabilities.

Figure 3.4. AI numeracy performance in 2016 and 2021, by question difficulty

Percentage share of numeracy questions that AI can answer correctly according to a simple majority of experts; measures use Yes/No-ratings, Maybe omitted



Source: Adapted from (Elliott, 2017^[2]), *Computers and the Future of Skill Demand*, <https://doi.org/10.1787/9789264284395-en>, and (OECD, 2023^[3]), *Is Education Losing the Race with Technology?*, <https://doi.org/10.1787/73105f99-en>.

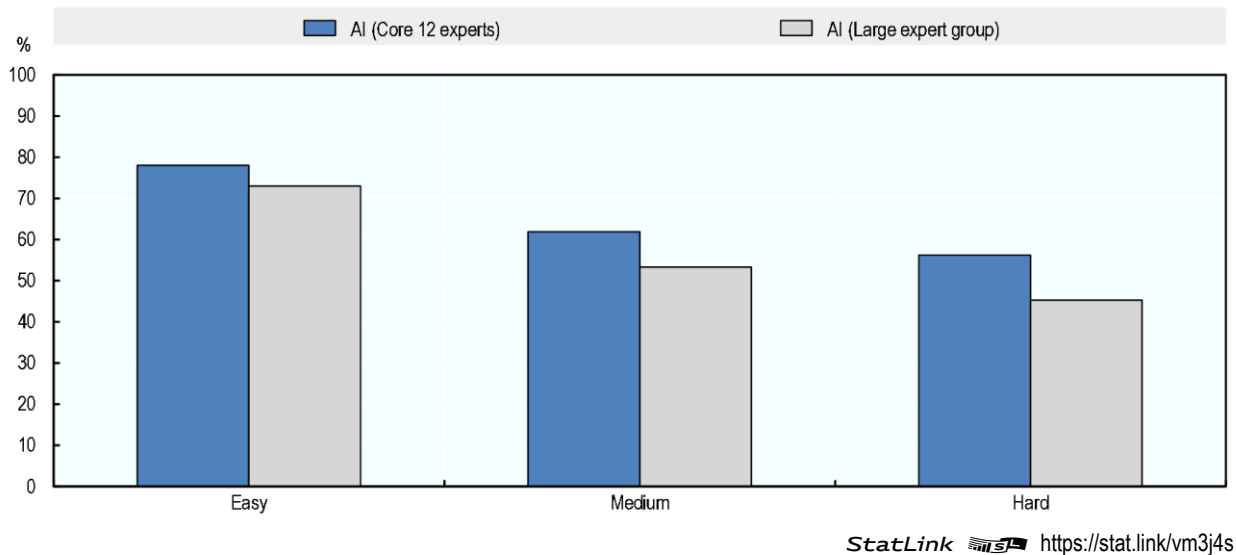
StatLink  <https://stat.link/2ryum7>

Figure 3.5 presents results from the assessment using PISA science questions, distinguishing between the 12 core experts and the larger expert group. It shows that both groups of experts have similar high confidence in AI solving easier questions, and lower confidence for more difficult ones. Overall, AI is expected to solve easy questions (Levels 1 and 2) at 78% confidence in the core expert group and at 73% confidence in the large expert group, which decreases to 62% and 53%, respectively, for questions of medium difficulty (Levels 3 and 4), reaching 56% and 45%, respectively, for hard questions (Levels 5 and 6).

The science questions consist mostly of text inputs. Therefore, the similarity of the results to those obtained with the PIAAC literacy test is not surprising. It reflects the strong performance of NLP systems in question-answering and text generation. That both the small- and the large-scale assessments produce similarly high ratings in this domain suggests that both the behavioural and the mathematical approaches to collecting expert judgement are effective in obtaining plausible evaluations from experts.

Figure 3.5. Predicted AI performance on PISA science questions in 2022 by core experts and larger expert group, by question difficulty

Average of experts' confidence in AI solving PISA science questions



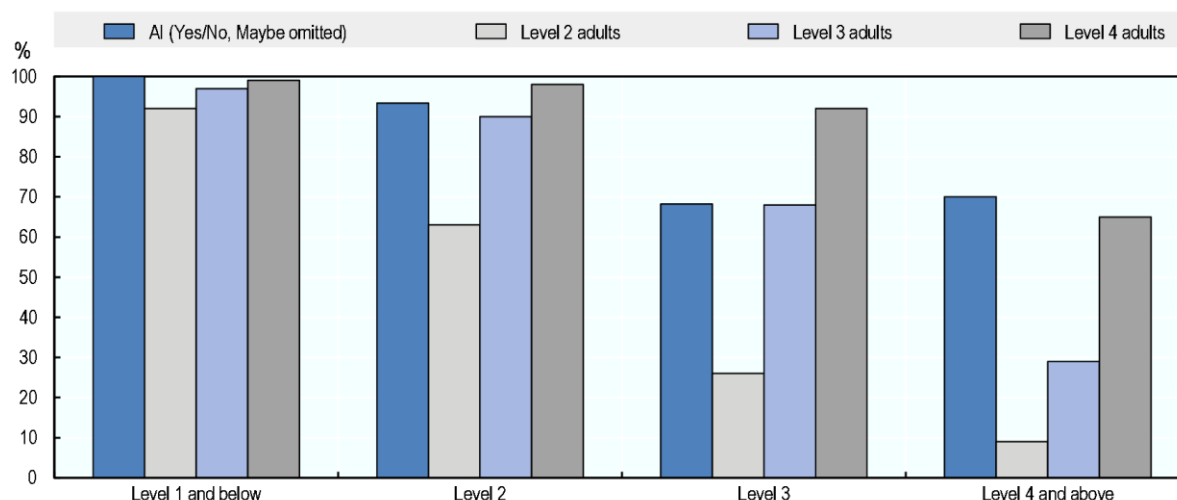
Comparison with human performance

As discussed above, the use of education tests for measuring AI capabilities offers the benefit of detailed comparisons to human performance. This provides insights into the potential impact of AI on essential skills used in educational and work settings. Tests like PIAAC go a step further in linking performance scores of respondents to various socio-economic and demographic characteristics. This allows, for example, for nuanced analyses of skill performance across countries, occupations, education levels or age groups.

Figure 3.6 illustrates how AI and human capabilities compare in literacy. The assessed AI performance by experts is compared to three proficiency levels of adult respondents. Adults at each proficiency level are expected to complete successfully 67% of the questions at that level. They have higher probability of success at easier questions, and lower chances to answer harder questions. The figure shows that expected AI performance resembles that of Level 3 adults. That is, AI is expected to solve about two-thirds of the Level 3 questions and almost all Level 1 and 2 questions. At Level 4, expected performance is actually closer to that of Level 4 adults, at 70%. However, this latter result should be interpreted with caution due to the small number of questions at that level.

Figure 3.6. Literacy performance of AI and adults of different proficiency

Share of literacy questions that AI can answer correctly according to the majority of experts compared to the probability of successfully completing items of adults at different proficiency levels



Source: Adapted from (OECD, 2023^[18]), *Is Education Losing the Race with Technology?*, Figure 4.6, <https://doi.org/10.1787/73105f99-en>.

PIAAC data show that most adults have literacy skills below Level 3. Across the OECD countries that participated in PIAAC, on average, 35% of adults are proficient at Level 3 and 54% score below this level; only 10% of adults perform better than Level 3 in literacy (OECD, 2019, p. 44^[11]). This suggests that AI can potentially outperform a large proportion of the population on the PIAAC literacy test.

Quality of AI measures

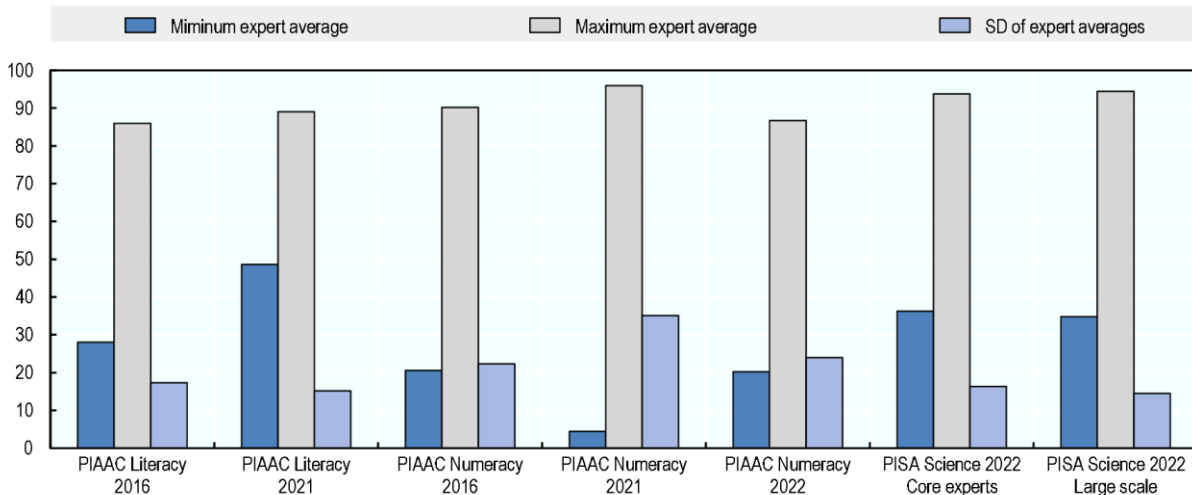
Disagreement among experts

If experts strongly diverge in their ratings, the assessment instrument likely lacks objective and clear criteria for rating, necessary for ensuring consistent results. In other words, the assessment instrument would not be reliable. This section looks at the diversity in experts' ratings, as it provides insights into inter-rater reliability.


Figure 3.7 shows the minimum and maximum average expert rating, as well as the standard deviation (SD) in these averages across the assessments. The average expert ratings are the means of each expert's ratings within an assessment. The figure shows that the highest variability in experts' overall judgements is in the numeracy domain (SD of 22.3 in 2016, 35.1 in 2021 and 24.0 in 2022), while SDs in literacy and science vary between 14.5 and 17.3. This reflects the strong disagreement in opinions in numeracy. In the first assessment round of the follow-up study with PIAAC, experts were uncertain how to interpret the scope of the numeracy tasks that an AI is supposed to master. Some assumed narrow tasks, while others focused on the entire range of tasks contained in the numeracy test. The result was two groups of experts with opposing opinions. Specifying the scope of tasks and clarifying the instructions for rating in the second round of the numeracy assessment resulted in agreement in the group discussion. However, there was still considerable variability in numerical ratings (OECD, 2023^[3]).

Figure 3.7. Divergence in experts' evaluations in different assessments

Minimum and maximum average expert rating and standard deviation of average expert ratings



Source: Adapted from (Elliott, 2017^[2]), *Computers and the Future of Skill Demand*, <https://doi.org/10.1787/9789264284395-en>, and (OECD, 2023^[18]), *Is Education Losing the Race with Technology?*, <https://doi.org/10.1787/73105f99-en>.

StatLink  <https://stat.link/zou86t>

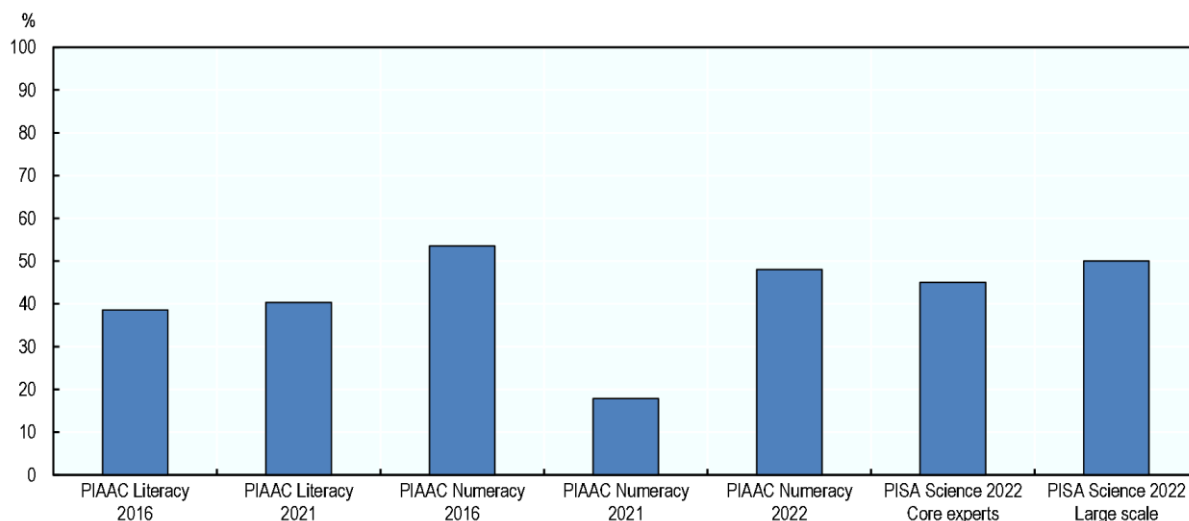
Uncertainty among experts

The degree to which experts are certain in their evaluations is instructive for the validity of measures. A high level of uncertainty among experts would suggest that the resulting indicators may not measure what they intend to measure. This would call for refining of assessment methodologies and the evaluation process.

Figure 3.8 shows the share of questions in each assessment that receive at least 20% of uncertain ratings. Uncertain ratings include the Maybe- and Don't know-categories. In the study with PISA, which uses a continuous scale for assessing experts' confidence, uncertainty is defined as confidence ratings in the 40-60% range. The share of questions receiving more than 20% of uncertain ratings is high in all assessments (between 39% and 54%). One exception is the first round of the follow-up numeracy assessment (18%). The 11 experts here expressed more certainty in their evaluations but had more opposing views on AI numeracy capabilities. Overall, the observations on experts' uncertainty show that obtaining valid ratings from experts is hard. Experts are not always knowledgeable about AI's potential to solve concrete tasks.

Figure 3.8. Share of questions that receive more than 20% of uncertain ratings in different assessments

Share of questions receiving at least 20% of Maybe-, Don't know-ratings or ratings within the 40-60% certainty range on PISA science questions



Source: Adapted from (Elliott, 2017^[2]), *Computers and the Future of Skill Demand*, <https://doi.org/10.1787/9789264284395-en>, and (OECD, 2023^[3]), *Is Education Losing the Race with Technology?*, <https://doi.org/10.1787/73105f99-en>.

StatLink  <https://stat.link/gno8v3>

Testing AI directly

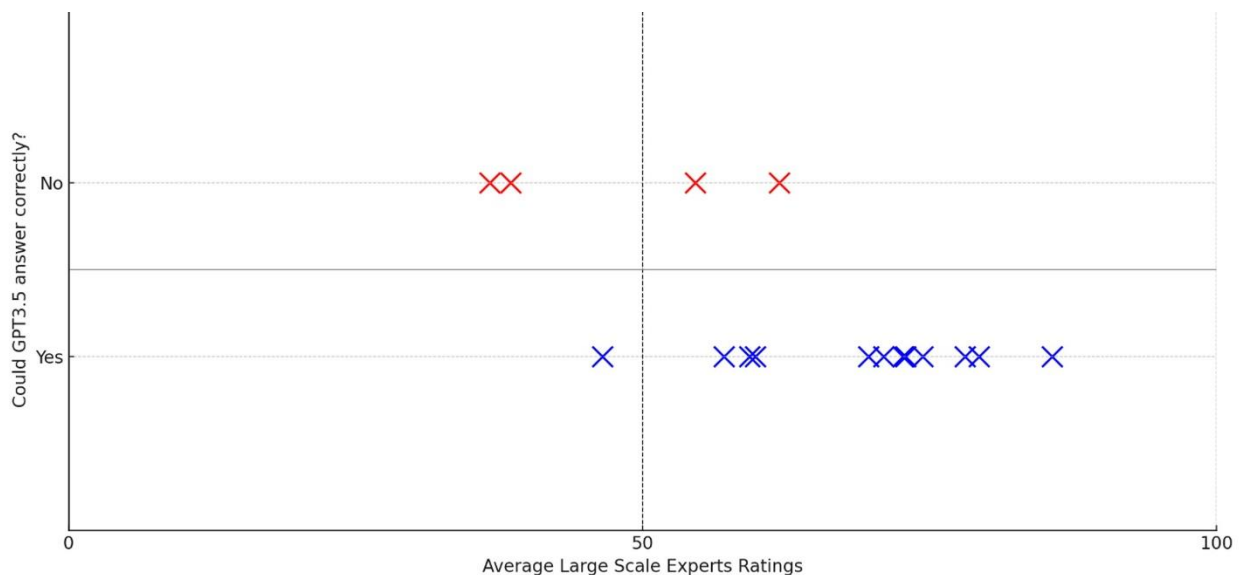
An effective way to evaluate the validity of the AI measures obtained from experts is to compare them to actual performance of state-of-the-art AI systems. In 2023, the project commissioned AI researchers to assess the performance of GPT-3.5 and GPT-4 on PISA reading, mathematics and science questions (OECD, 2023^[18]). Sixteen of the science questions were also evaluated by the experts. In the following, the expert ratings on these questions are compared to GPT-3.5 performance (Ye et al., 2023^[19]). This model was released in March 2022, before the assessment took place, and was an integral part of the first version of ChatGPT, released in November 2022.


GPT-3.5 performed well on easier questions, and less so on harder ones. It solved all of the test questions at Levels 1 and 2 and fewer questions at higher levels of difficulty. Overall, it could solve 12 of the 16 items. Figure 3.9 compares these results to the confidence ratings of experts. All questions that GPT-3.5 could solve have received high confidence ratings by experts, except for one. This latter question received an average rating of 47%, indicating uncertainty among experts regarding AI performance. Out of the four questions that GPT-3.5 could not solve, two were incorrectly rated as likely solvable. However, these ratings are again closer to uncertainty (55% and 62%).

Overall, these findings show that experts, although uncertain in many matters, can correctly assess the capabilities of current state-of-the-art of AI technology.

Figure 3.9. Experts' ratings of AI and GPT-3.5 performance on PISA science questions

Means of experts' ratings on 16 PISA science questions, by correct and incorrect response by GPT-3.5



StatLink  <https://stat.link/f0bmed>

Lessons learnt

The results of the online assessments and, above all, the discussions with experts revealed a number of challenges in using education tests to collect expert judgement on AI. The project team worked together with the experts to develop methodological solutions. This section outlines the major takeaways from this work.

Quantitative disagreement, qualitative agreement

It proved difficult to obtain coherent expert ratings on AI's capability to solve the education tests. Part of this difficulty related to differences in how experts interpreted the scope of the test questions that a hypothetical system is supposed to master. Discussions with experts showed that some were inclined to rate AI on subsets of similar test questions, similar to how AI systems are typically trained and evaluated in practice. Other experts evaluated current systems' capability to tackle all test questions in a domain at once. Overall, experts agreed that developing tailored solutions for narrow tasks is easier than developing general systems that can tackle all types of questions, including similar questions that are not part of the test. The project team attempted to specify the assumed scope of the test in order to translate this qualitative agreement into coherent numerical ratings.

The proposed new instructions for rating were tested in a second round of the follow-up study with the PIAAC numeracy test and in the study with PISA. The ratings obtained with this method from the assessment with PISA did not diverge strongly, as shown in Figure 3.7. However, the new framing has only partially reduced disagreement in the PIAAC numeracy assessment. That is, compared to the 11 experts who completed the first round of the follow-up numeracy assessment, the 4 experts in mathematical reasoning who re-assessed numeracy with the new framing showed lower – but still substantial – variation in their ratings.

The discussion with the four experts showed that they clearly interpreted the scope of tasks that a hypothetical system is supposed to solve from the instructions. However, they differed in the time frame

which they used for the evaluation. Experts were instructed to consider a hypothetical engineering effort to develop a system for the test using state-of-the-art AI techniques. The experts with the highest ratings argued that, given rapid advancements in the field, such an effort would produce the desired results within less than one year.³ By contrast, the expert with the lowest ratings focused on the current state of AI systems, which were not able to solve the numeracy test at the time. However, he agreed that systems will likely reach this stage within a year.

Overall, the methodological changes introduced in the rating exercise have increased clarity and consensus about AI capabilities. However, there is still need for improvement. The instructions need to reflect the fast pace of AI progress by using shorter time frames for rating. This would enable more precise evaluations of the state of the art of AI capabilities.

Literacy easier to rate than numeracy

Experts seemed at ease considering the application of NLP systems on the PIAAC literacy test. During group discussions, they noted that the literacy questions are similar to real-world tasks addressed by existing applications. In addition, benchmark tests used for evaluating NLP systems, such as the Stanford Question Answering Dataset (SQuAD) (Rajpurkar, Jia and Liang, 2018_[20]), often contain similar problems and tasks. Therefore, experts saw PIAAC as an appropriate tool for evaluating potential AI performance in language processing.

By contrast, the 11 experts who first rated AI in numeracy in the follow-up study described the exercise as less straightforward than the literacy assessment. They saw the numeracy questions as more distant from problems typically addressed by AI research. Until 2021, AI research had paid less attention to mathematical reasoning of AI because of its relatively lower applicability and commercial use. In addition, the mathematical tasks typically addressed in research – automated theorem proving and math word problems (i.e. quantitative problems stated in text), among others – are different from the ones in PIAAC. This made it challenging for experts to rate potential AI performance on the test. For the four experts in mathematical reasoning, the evaluation was easier due to their better understanding of the domain.

This suggests that expert knowledge elicitation on AI capabilities is more feasible in domains that are an established application domain in AI research. In less prominent or novel domains, experts have more limited information on research results and existing systems, unless they have specialised knowledge in the relevant domain.

More experts do not add value to the results

The study with PISA showed that the behavioural and mathematical approaches for expert knowledge elicitation produce similar results on AI capabilities. Expert ratings obtained with the mathematical approach also show similar variability compared to the ratings of the core experts. In addition, the similarity of these ratings with the performance of the contemporaneous GPT-3.5 system on the PISA science questions provides some evidence of their validity.

However, the advantages of this approach – obtaining robust measures that reflect the opinions of a large number of experts – do not outweigh its disadvantages. As described in Chapter 2, recruiting a large sample of experts proved challenging. Among the 189 computer scientists who were contacted by the project team, 63 expressed interest in participating in the survey, and only 33 actually completed it. Monetary incentives played a strong role in this process, suggesting that a repeated large-scale assessment of AI capabilities will be costly to implement.

As a result of these explorations, the project has chosen to rely on input from small groups of familiar experts for future activities involving expert knowledge elicitation. The study with PISA confirmed the robustness of this approach to the use of many experts.

The way forward

The exploratory studies using PIAAC and PISA have important implications for the methodology of the project. They identified limits to obtaining robust measures of AI capabilities by surveying experts. Consensus evaluations are hard to obtain, especially in domains that are not the centre of current research. This was the case for AI quantitative reasoning at the time of the PIAAC numeracy assessment. In addition, the assessment is time-consuming, both for the experts who need to invest several hours to provide ratings and participate in discussions, and for the project staff who devoted substantial time to recruit and engage experts. This led the project to test out the use of available direct measures of AI, which are discussed in Chapters 6 to 8.

However, expert judgement remains an indispensable part of the methodology. It is needed for reviewing, selecting and interpreting existing measures of AI. Measures obtained from expert evaluations can also complement the overall assessment framework in areas in which results from direct assessments of AI systems are lacking. For example, research interest and investment in automating particular tasks may be limited because their practical applicability and economic benefits may not be immediately clear. Other tasks may receive less research attention because they are still clearly out of the reach of current state-of-the-art technologies. Expert judgement can help fill such gaps by providing information on how far AI is from performing such tasks. In this way, the approach can contribute to a comprehensive assessment of AI capabilities across a wide range of human skills.

References


- Bolger, G. (ed.) (1992), *Perspectives on Expertise in the Aggregation of Judgments*, Plenum. [15]
- Bubeck, S. et al. (2023), “Sparks of Artificial General Intelligence: Early experiments with GPT-4”. [7]
- Clark, P. and O. Etzioni (2016), “My Computer Is an Honor Student — but How Intelligent Is It? Standardized Tests as a Measure of AI”, *AI Magazine*, Vol. 37/1, pp. 5-12, <https://doi.org/10.1609/aimag.v37i1.2636>. [9]
- Drori, I. et al. (2021), “A Neural Network Solves, Explains, and Generates University Math Problems by Program Synthesis and Few-Shot Learning at Human Level”, <https://doi.org/10.1073/pnas.2123433119>. [24]
- 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]
- Frieder, S. et al. (2023), “Mathematical Capabilities of ChatGPT”. [8]
- Hendrycks, D. et al. (2020), “Measuring Massive Multitask Language Understanding”. [5]
- Hendrycks, D. et al. (2021), “Measuring Mathematical Problem Solving With the MATH Dataset”. [22]
- Lewkowycz, A. et al. (2022), “Solving Quantitative Reasoning Problems with Language Models”. [23]
- 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>. [3]
- OECD (2023), *Putting AI to the test: How does the performance of GPT and 15-year-old students in PISA compare?*, OECD Education Spotlights, No. 6, OECD Publishing, Paris, <https://doi.org/10.1787/2c297e0b-en>. [18]
- OECD (2019), *PISA 2018 Results (Volume I): What Students Know and Can Do*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/5f07c754-en>. [4]
- OECD (2019), *Skills Matter: Additional Results from the Survey of Adult Skills*, OECD Skills Studies, OECD Publishing, Paris, <https://doi.org/10.1787/1f029d8f-en>. [1]
- OECD (2016), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264266490-en>. [13]
- OECD (2016), “PISA 2015 test items”, in *PISA 2015 Results (Volume I): Excellence and Equity in Education*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264266490-15-en>. [21]
- OECD (2013), *The Survey of Adult Skills: Reader’s Companion*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264204027-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]
- OECD (2009), *Take the Test: Sample Questions from OECD’s PISA Assessments*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264050815-en>. [12]

- OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264040014-en>. [14]
- OpenAI (2023), *GPT-4 Technical Report*, <https://cdn.openai.com/papers/gpt-4.pdf>. [6]
- Radford, A. et al. (2018), *Improving Language Understanding by Generative Pre-Training*, https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 1 February 2023). [17]
- Rajpurkar, P., R. Jia and P. Liang (2018), "Know What You Don't Know: Unanswerable Questions for SQuAD". [20]
- Tversky, A. and D. Kahneman (1974), "Judgment under Uncertainty: Heuristics and Biases", *Science*, Vol. 185/4157, pp. 1124-1131, <https://doi.org/10.1126/science.185.4157.1124>. [16]
- Ye, J. et al. (2023), "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models". [19]

Annex 3.A. Analyses of the PIAAC and PISA studies using an alternative approach

Annex Table 3.A.1. List of online figures for Chapter 3

Figure Number	Figure Title
Figure A3.1	AI literacy performance in 2016 and 2021, following the “average” approach
Figure A3.2	AI numeracy performance in 2016 and 2021/22, following the “average” approach
Figure A3.3	AI performance on PISA science questions in 2022, following the “majority” approach

StatLink  <https://stat.link/glv4ed>

Notes

¹ See Note 2 in Chapter 1 of this volume.

² All items used in this AI assessment are sourced from the publicly released examples of the PISA 2006 and 2015 editions (OECD, 2009^[12]; OECD, 2016^[21]). The publicly released items contain limited information about students' performance on the questions. However, they include information on question difficulty and the sub-skills involved.

³ Prior to the assessment, which took place September 2022, the field of mathematical reasoning of AI has taken major steps. In 2021, the MATH dataset, a leading benchmark for mathematical reasoning, was released (Hendrycks et al., 2021^[22]). Between 2021 and 2022, several large language models fine-tuned for quantitative problems were launched (Lewkowycz et al., 2022^[23]; Drori et al., 2021^[24]). In addition, major AI labs were close to developing multimodal systems that can process both images and text. Experts referred to these developments, reflecting on the likelihood of AI solving the numeracy test in the near future.

4 Occupational tests

Mila Staneva, OECD

Britta Rüschoff, FOM University of Applied Sciences for Economics and Management

Phillip L. Ackerman, Georgia Institute of Technology

This chapter describes performance tests on occupational tasks stemming from occupation certification and licensure examinations. It discusses use of such examination tasks for collecting expert judgement on artificial intelligence (AI) and robotics performance. The chapter describes 13 example tasks from six occupations selected for an explanatory assessment of AI and robotics. The tasks were chosen from final examinations in German vocational education and training, as well as certification and licensure exams used in the United States. The chapter describes the development and administration of such tests, their types and formats, as well as procedures to ensure their content validity. It concludes with discussing methodological steps towards a comprehensive and robust approach for studying the capabilities of AI and robotics and their impact on occupations.

Understanding the implications of evolving technologies for education and employment requires an evaluation of artificial intelligence (AI) and robotics across a wide range of skills used in the workplace. Next to key cognitive skills, such as literacy and numeracy, the workplace involves various occupation-specific technical skills and domain-specific professional knowledge. Accordingly, a battery of different instruments for measuring AI and robotics capabilities is needed.

As a complement to the education tests discussed in Chapter 3, this chapter explores the use of complex occupational tasks for collecting expert ratings on AI and robotics capabilities. These tests stem from licensing and certification examinations for different occupations. They include typical (hands-on) tasks in the occupation, such as a product designer creating a design for a new container lid, or a cosmetologist performing a manicure. Experts' ratings of AI and robotics' performance on such tasks can provide valuable insights into the readiness of these technologies for real-world applications and their potential to replace or support workers in their jobs.

This chapter discusses the usefulness of complex occupational tasks for evaluating AI and robotics. Tasks from two sources are presented – final examinations in German vocational education and training (VET) as well as certification and licensure exams used in the United States. The chapter describes the development and administration of such tests, their types and formats, and procedures to ensure content validity. It then outlines 13 tasks from six occupations selected for an exploratory evaluation of AI and robotics' performance using expert judgement. The chapter concludes with discussing further steps towards an approach for assessing AI and robotics capabilities across a wide range of occupational tasks and analysing how progress in capabilities may change these tasks.

Rationale for collecting expert judgement on AI with complex occupational tasks

Like other human tests used to evaluate AI in this project, examinations that certify workers for specific occupations have a variety of advantages. They offer computer experts standardised and objective evaluation criteria for rating AI and robotics' performance. This allows for consistent ratings across different expert groups and across time. They provide precise, contextualised and granular descriptions of the tasks. This allows experts to make exact judgements of AI's potential performance on the task (and parts of the task) and, thus, improves reliability across experts. Moreover, occupation certification and licensure tests provide a way to compare AI and robotics performance to human performance. Where data on the performance of human test takers are missing, the passing score on the exam can indicate whether machines satisfy the minimum skills requirements that workers must fulfil to enter occupations.

An additional advantage of certification and licensure examinations is related to the use of real-world tasks and scenarios that are typical for occupations. These tasks are action-oriented and practical, drawing on observable task-related behaviour (e.g. crafting a certain product). The practical relevance is provided in several ways. For example, test items are selected on the basis of careful job analysis that determines the most important and most frequently applied tasks in occupations (Johnston et al., 2014^[1]). People working in the profession, so-called “subject matter experts”, and/or industry representatives are often involved in the test design. Examinations are aligned to the framework curricula of training programmes, which, for their part, mirror industry standards and best practices (Rüschoff, 2019^[2]).

This practical orientation of occupation certification and licensure tests distinguishes them from other types of tests that rely on broad, underlying characteristics of applicants to predict job performance. The latter tests seek to assess abstract constructs, such as general intelligence, broad content abilities (e.g. verbal, spatial, numerical abilities) or narrower abilities (e.g. perceptual speed, psychomotor abilities), based on evidence that these traits manifest in various behaviours, including how one performs job-related tasks. For example, individuals who score high on a general mental ability test are more likely to successfully perform tasks involving complex problem-solving in real-life situations. The tasks included in the ability test can be thus indicative for an observable behaviour of interest. However, these tasks are indicators of an

abstract, underlying psychological construct rather than direct demonstrations of the behaviour. They are meaningful insofar as they are correlated with the behaviour of interest.

By contrast, the occupational tasks included in certification and licensure examinations have considerable meaning on their own. They assess concrete professional behaviours with immediate relevance to exercise of the profession. The resulting measures rely far less on theoretical assumptions regarding underlying psychological constructs. This poses an advantage for assessing AI and robotics since a theoretical link between concrete performance on a task and broad underlying abilities cannot be assumed for machines. In other words, high performance on a task cannot be attributed to a general ability that would enable high performance on another, different task (OECD, 2023^[3]).

As shown in Chapter 3, the use of tests targeting broad underlying foundation skills, such as literacy and numeracy, posed challenges to the assessment. Experts diverged in their ratings of AI performance on these tests because they were uncertain how general the computer capabilities being assessed are supposed to be. They argued that the capacity of systems to solve one test task does not presuppose high performance on other task types and formats. To make precise judgements, the experts thus needed clarification on the generality of the underlying capabilities being evaluated. However, defining generality is not trivial. It requires a specification of all tasks that a system is supposed to master, within and beyond the test.

Testing AI and robotics on concrete occupational tasks should mitigate this problem. The measures of this type of task performance are also narrow in their generalisability. The reason is that occupational tasks are complex, involving multiple actions and requiring different capabilities, which makes it hard to attribute success on one task to other tasks or unknown contexts. Still, performance tests from occupation entry examinations would show whether a machine can or cannot complete typical tasks in an occupation.

Occupational tasks from certification and licensure examinations

Many occupations require passing an exam that establishes whether candidates demonstrate the requisite knowledge, skills and abilities to engage in practice. Additional minimum entry requirements are often in place, such as qualifications, experience or medical record. These occupational entry regulations serve the purpose of protecting the public from unqualified practitioners and ensuring good quality services and products through standardising the skills of their providers (Koumenta and Pagliero, 2017^[4]). Such considerations are especially strong for occupations that are of particular interest for the public good.

Occupational entry regulations can take different forms. Licensure is the strictest form of regulation. It grants those who can demonstrate the specified level of competence the legal right to exercise protected activities. Persons without a licence cannot practice the occupation. By contrast, certification provides a legally protected title that indicates a minimum competence for an occupation. Those who do not hold the certificate are not legally restricted from carrying out tasks covered by the occupation. While licensing is overseen by government or state authorities (or appointed regulators), certification programs can also be developed by professional associations, chambers of industry or other membership organisations.

Occupation licensure and certification practices vary across countries and occupations. In the European Union, 43% of workers held a certificate or a licence in 2015 (Koumenta and Pagliero, 2017^[4]). The proportion of licensed workers was highest in Germany, at 33%, and lowest in Denmark, at 14%. The proportion of certified workers varied between 36% in Germany and 9% in Finland (Koumenta and Pagliero, 2017^[4]). In the United States, 28% of the workforce was licensed or certified in 2013, with big differences among the states (Kleiner and Vorotnikov, 2017^[5]). In general, occupational licensing and certification are more typical for teaching professionals, health and social workers, and plant and machine operators, and is less common for managers, in wholesale or retail services, agriculture or elementary occupations (Koumenta and Pagliero, 2017^[4]). However, the same occupation can be subject to very

different entry regulations in different jurisdictions. In the United States, for example, fewer than 60 occupations were regulated in all 50 states. Meanwhile, more than 1 000 occupations were regulated in at least one state (Kleiner and Vorotnikov, 2017^[5]).

To account for possible country differences in examination practices, this study uses occupational tests from two countries – Germany and the United States.

German VET assessments

Germany has 324 vocational occupations that are state-recognised under the Vocational Training Act (BBiG) or the Crafts Code (HwO) (BIBB, 2022^[6]). Other state-regulated occupations, such as in the medical field, are covered in special laws (e.g. Nursing Act, Geriatric Care Act). The dual VET system provides entry into most vocational occupations. This study refers primarily to dual VET since it is the most common training model that qualifies workers for occupations in Germany.

In dual VET, apprentices acquire theoretical knowledge by attending a vocational school and receive practical training by working in a company. Apprentices sign a contract with the company and receive remuneration for their work. In-company training is regulated by the Vocational Training Act and by the training regulations of the occupations (*Ausbildungsordnungen*). School-based vocational education is regulated by framework curricula (*Rahmenlehrpläne*). The training regulations and the framework curricula provide national standards regarding training content, training facilities, trainers and examinations (Cedefop, 2020^[7]).

At the end of their training, apprentices complete a final exam to obtain a certificate. Final examinations in (dual) VET are regulated by the Vocational Training Act (BBiG, §37 – §50) or the Crafts Code (HwO, §31 – §40a). They are organised by the respective chambers for each occupation. The chambers appoint examination boards and conduct the examinations, which are developed in accordance with the relevant regulatory instruments – training regulations and the framework curriculum (OECD, 2021^[8]).

Examinations are aligned with the curricula to reflect all relevant domains in the occupation. Examples of professional behaviour in the respective occupation are provided for each examination domain. These examples form the basis for the examination tasks. The tasks are commonly classified by content and the competences they aim to assess (Badura, 2015^[9]).

Different test developers follow different competence models when developing tasks. For example, AKA (*Aufgabenstelle für kaufmännische Abschluss- und Zwischenprüfungen*)¹, which develops examinations for commercial occupations, classifies tasks according to the content domains planning, execution and evaluation of results and according to whether tasks assess knowledge or skills and abilities (Badura, 2015^[9]). The classifications used are not always aligned with international skills taxonomies commonly considered in research and policy.

Examination development offices usually develop the tasks. For example, PAL (*Prüfungsaufgaben- und Lehrmittelentwicklungsstelle*)² develops exams for industrial and technical professions, ZPA (*Zentralstelle für Prüfungsaufgaben*)³ or AKA cover commercial professions, and the ZFA (*Zentral-Fachausschuss Berufsbildung Druck und Medien*)⁴ provides examinations in the field of printing and media. However, other entities may also develop tasks. Examples include the examination board appointed by the Chamber of Commerce and Industry (IHK) or a task development committee appointed by the respective chambers.

The tasks are developed in close co-operation with the industry to ensure their content validity i.e. that the task content fully represents the requirements and content of the occupation. Another instrument for establishing content validity are examination catalogues or grids. An examination grid indicates the proportional distribution of the tasks across examination domains. It is based on the specifications in the regulatory instruments. The grid is intended to ensure the examination covers all content domains relevant in an occupation and to give domains a correct weighting. In addition to the validity of the examinations,

the development offices commonly keep a record of the reliability (i.e. whether a test produces similar results under consistent conditions) and discriminatory power of the tests (i.e. whether a test can distinguish between two or more groups being assessed).

US licensing and certification exams

In the United States, occupational credentialing and licensure are largely decentralised. State governments generally enact the laws regulating occupational licensing. Some states embed these requirements directly in the statute authorising creation of the licence. Other states authorise their agencies or state-sponsored independent boards to develop licensure requirements. Often, occupation entry requirements combine both – statute and regulations set by a designated agency or board (NCSL, 2022_[10]).

As a result, there are significant differences in licensing requirements across the states (NCSL, 2022_[10]). For example, the State of Georgia has 180 different occupational licences. Many are different “levels” or types of the same occupation, such as separate certifications for nine different nurse licences. In comparison, the State of California has 357 different occupational licences (US Department of Labor, 2021_[11]). In some instances, reciprocity agreements make it easier for licensees in one state to be licensed in another.

The states typically delegate implementation of occupational entry regulations to professional associations. The latter usually form one or more intermediary agencies to assume responsibility for development and validation of examinations (e.g. National Council of Architectural Registration Boards, American Board of Dental Examiners). In other instances, agencies or organisations that award credentials develop examinations for credentials on their own (e.g. National Commission on Certification of Physician Assistants (Buckendahl, 2017_[12]). In addition, third-party auditors such as the National Council of Certifying Agencies and the American National Standards Institute have developed formal standards for evaluating credentialing programmes (Johnston et al., 2014_[1]).

Test development for certification or licensure is typically a function of formal or informal job analyses, the engagement of subject matter experts, and attention to a body of reference materials for formulating test questions and determining the scope of the examinations. Job analysis involves different methods to identify the mainstream activities of the profession. In addition, it collects information about the knowledge, skills and abilities (KSAs) needed for performing these activities. This information is typically collected from surveys of subject matter experts. Other sources of information include course outlines, laws, textbooks or other curricular materials. The apparent goal of these design characteristics is to improve the content validity of the test rather than construct validity (i.e. Does the test measure the construct it intends to measure?) or criterion-related validity (i.e. Do test results correlate with results from other tests measuring the construct?).

The results of the job analysis are used to develop a so-called test blueprint. This document serves as the basis for developing concrete examination tasks by specifying the most important characteristics of the test. These are the behaviours or KSAs to be assessed, but also other features, such as the emphasis given to different domains, the item format or the difficulty of the test. In addition to providing item developers with direction, blueprints help ensure continuity in test content and difficulty over time and serve as the basis for item classification and revision. They can also inform educators and test takers of test content and assist them in their test-preparation efforts.

Formats of performance tests and grading

The tasks included in certification and licensure examinations can be written, oral or practical (hands-on procedural) demonstrations of knowledge and skills. Written tasks can have different formats (e.g. open-answer, multiple-choice, fill-in-the-blank questions). They are often applied, in the sense that they cover typical professional activities, such as writing a business letter in commercial professions or

documenting the manufacture of a product in technical professions. Examples for oral tasks are presentations, conversation simulations or discussing a completed assignment. Procedural demonstrations typically include crafting a product or carrying out a typical activity for the profession (BIBB, 2013^[13]).

The grading of occupation examinations varies widely. Licensure examinations often use a binary pass/fail grading since they aim at determining whether the examinee has a “minimum competence” for the profession. Different cut scores can determine whether an examinee passes or fails. Some examination procedures use absolute scores (e.g. 60% correct). Other examinations use “scaled scores”. These are essentially norm-referenced measures where passing depends on the individual’s percentile rank among other test takers. Classifying an examinee as qualified/non-qualified is thus determined partly by how others perform in the examination.

By contrast, certification exams usually use continuous-graded scales or categorical ratings, such as “novice, apprentice, journeyman and expert” (Hambrick and Hoffman, 2016^[14]). This grading is more suitable for evaluating computer performance on occupational tasks and comparing it to human performance. However, even when examinations designate individual examinees as qualified or not qualified, there is an underlying continuous scale upon which individuals perform. Thus, the application of occupation tests for evaluating AI and robotics should not be limited to a dichotomous evaluation.

Selection of occupations and examination tasks

The project selected 13 example tasks from six occupations – amid a larger pool of identified occupations – to explore their use for assessing AI and robotics capabilities. The aim was to select diverse examination tasks representing some important elements of reasoning, language and sensory-motor capabilities. In addition, the tasks covered different occupations and working contexts and had different levels of complexity to explore how these different aspects relate to the collection of expert judgement.

Several considerations guided the selection of occupations:

First, occupations were sampled across the broad categories of the International Standard Classification of Occupations (ISCO) (ILO, 2012^[15]). ISCO divides occupations into ten major groups, which are further divided into smaller subgroups. The categorisation of occupations into broad groups depends on the skill level and the education required for occupations. Occupations from five of the ten major groups were selected to cover different levels of the occupational hierarchy (see Table 4.1).

Table 4.1. Selected occupations

Occupation	ISCO occupational domain	ISCO 4-digit code	NACE Industry sector
Specialist in metal technology	7 Craft and related trades workers	7212 - 7215	C Manufacturing
Cosmetologist	5 Service and sales workers	5141	S Other service activities
Office management assistant	4 Clerical support workers	4110	N Administrative and support service activities
Dental assistant	3 Technicians and associate professionals	3251	Q Human health and social work activities
Nursing professional	2 Professionals	2221	Q Human health and social work activities
Technical product designer	2 Professionals	2163	M Professional, scientific and technical activities

Note: The International Standard Classification of Occupations (ISCO) has ten broad groups according to required skill level and qualification: 1 Managers; 2 Professionals; 3 Technicians and associate professionals; 4 Clerical support workers; 5 Service and sales workers; 6 Skilled agricultural, forestry and fishery workers; 7 Craft-related trades workers; 8 Plant and machine operators, and assemblers; 9 Elementary occupations; 0 Armed forces occupation (ILO, 2012^[15]). The Statistical Classification of Economic Activities in the European Community (NACE) is an international taxonomy of industries that distinguishes 21 major industrial sectors (European Commission, 2008^[16]).

Box 4.1. Direct assessment of large language models on written professional certification tests

A number of studies applied GPT (Generative Pre-trained Transformer) language models on written professional certification exams to study their content-specific capabilities. For example, Noever and Ciolino (2023^[17]) assessed the performance of GPT-3 and GPT-3.5 on a test dataset containing 1 149 professional certifications. They showed that GPT-3 could solve more than 70% of the test questions on 39% of the exams. GPT-3.5 demonstrated better performance on many exams, particularly those for accountants, veterinarians, aviation inspectors, real estate appraisers, human resources professionals and financial planners.

Several studies evaluated the large language models in the domain of medicine. For example, Ali et al. (2023^[18]) tested GPT-3.5 and GPT 4 on a neurosurgery written examination. GPT-3.5 scored 73.4% and GPT-4 scored 83.4% on the test, compared to an average of 73.7% of human tests takers. Antaki et al. (2023^[19]) evaluated ChatGPT, based on GPT-3.5, in the domain of ophthalmology and found that the model had modest overall performance. By contrast, Lin et al. (2023^[20]) evaluated both GPT-3.5 and GPT-4 on ophthalmology written examination and found that GPT-4 performance (76.9%) exceeds both the performance of its predecessor (63.1%) and human performance (72.6%). Nori et al. (2023^[21]) tested GPT-4 on practice materials for the United States Medical Licensing Examination (USMLE). They show that the model exceeds the passing score of the test by 20 points and outperforms GPT-3.5 as well as models specifically developed for the medical field.

Other studies showed that GPT-4 passed US license exams in accounting (Eulerich et al., 2023^[22]) and law (Katz et al., 2023^[23]).

While these studies evaluate large language models on written exams, the exploratory study described in Chapter 5 asks experts to assess the state of the art in AI and robotics on mostly practical tasks from occupational examinations. As described in this chapter, the examination materials used in this study are diverse, testing capabilities such as vision, planning or dexterity.

Second, occupations from different industries were selected. Following the Statistical Classification of Economic Activities in the European Community, commonly referred to as NACE, five industries were covered: manufacturing; health; professional, scientific and technical activities; administrative and support service activities; and other service activities (European Commission, 2008^[16]).

Different examination materials were retrieved for each selected occupation (see Table 4.2). Due to the large scope of the examinations, only parts of the exams or single tasks were used. The criteria for selecting these materials aimed again at diversifying the set of examination tasks in order to test different possibilities for assessing AI and robotics performance on the job. First, materials from both US and German examinations were selected. Second, both practical and written tasks were used. The written tasks contained both open (e.g. writing an e-mail) and closed (e.g. fill-in-the-blank task) questions. The practical tasks ranged from shorter tasks to comprehensive work assignments that took examinees several hours to complete. Some materials contained specific instructions, including a list of the instruments available to the examinees and supplementary materials, such as technical drawings or plans. Other tasks were described briefly. This aimed to test how much information on each task experts need to reliably judge AI and robotics potential performance. Finally, tasks were chosen to cover a wide range of skills, from sensory-motor skills to reasoning and language and use of domain-specific knowledge. (Noever and Ciolino, 2023^[17])

Table 4.2. Selected occupational tasks

Occupation	Task content	Examples of skills required in the task	Task format and provider
Specialist in metal technology	Manufacture a functional assembly according to given specifications; assess and document whether the components of the product are dimensionally accurate in a measurement protocol.	Knowledge of tools and materials; technical knowledge of the process steps in manufacturing products; planning; ability to autonomously execute work orders according to technical instructions; subject-specific mathematical skills, e.g. reading units of measurement; general dexterity; general mathematical skills; spatial thinking/spatial imagination.	Practical task (DEU) ¹
	Use two pieces of sheet aluminium and filler rod to weld a Tee-joint in the horizontal position.	Ability to select and set up equipment correctly and safely; ability to select and use the right material (aluminium and filler rod); ability to weld according to specifications.	Practical task (US) ²
Cosmetologist	Perform chemical waving.	Knowledge of chemical waving supplies; ability to wrap hair; ability to place rod correctly throughout entire section; ability to understand and follow instructions.	Practical task (US) ³
	Perform a full manicure including a hand massage, remove excess cream from nails, and polish nails.	Knowledge of materials and tools; ability to use correct manicure techniques; ability to perform a hand massage; ability to understand and follow instructions; ability to clean workstation.	Practical task (US) ²
Office management assistant	Prepare an evaluation of the complaints based on available data; find several files saved in different folders and use spreadsheets from these files according to given specifications.	ICT skills; general literacy skills; general mathematical skills; ability to read tables; ability to perform spreadsheet calculations (e.g. application of formulas); analytical ability to translate data into meaningfully visualised diagrams; ability to analyse business data and to communicate the results of these analyses appropriately.	Practical task (DEU) ⁴
	Draft an e-mail to line manager explaining the results of the analysis of the complaints and proposing solutions.	ICT skills; general literacy skills; ability to write professional communication (e-mail); ability to formulate written communication (by e-mail) appropriate to the addressee; ability to handle and understand tables/data; ability to autonomously develop ideas for solutions based on data and professional knowledge; ability to communicate own ideas and proposed solutions in a comprehensible way (in writing).	Writing task (DEU) ⁴
	Create a flyer using specifications provided.	ICT skills; general literacy skills; ability to understand and follow instructions; familiarity with formatting techniques and practices; ability to use and maintain office equipment (copier).	Practical task (US) ²
Dental assistant	Indicate the different groups of teeth and their distribution in the deciduous and permanent dentition in the blank spaces of the table provided.	Knowledge of Latin dental terminology; know differences between deciduous and permanent dentition; general literacy skills.	Writing task (DEU) ⁵
	Name two ways of performing a sensitivity test.	Knowledge of dental terminology; knowledge of dental exam procedures; ability to identify correct dental procedure based on a given situation and exam purpose.	Writing task (DEU) ⁵
	Prepare instruments for autoclaving.	Knowledge of correct materials and equipment for autoclaving; knowledge of pre-cleaning, disinfection and sterilisation procedures according to federal guidelines; ability to apply the correct safety and sanitation procedures in preparing dental instruments; ability to understand and follow given instructions.	Practical task (US) ²
Nursing professional	Transfer a Cerebral Vascular Accident patient with right-side paralysis from bed to wheelchair and back to bed.	Knowledge of necessary equipment; ability to identify the patient; ability to introduce and explain the procedure; ability to use equipment and aseptic techniques properly and safely; ability to identify appropriate patient positioning, transfers and body alignment; ability to position patient in a wheelchair and bed.	Practical task (US) ²

Occupation	Task content	Examples of skills required in the task	Task format and provider
Technical product designer	Produce technical drawing proposals of modifications to a kitchen tool using Computer-Aided Design (CAD).	Ability to translate technical drawings and functional descriptions into work orders; planning skills; general literacy; analytical/mathematical ability to calculate and complete missing dimensions autonomously using known parameters; ability to make autonomous design decisions based on a functional description and known parameters; ability to produce different types of drawings (e.g. exploded view); basic technical/physical knowledge; ability to use CAD.	Practical task (DEU) ⁶
	Create a 3D solid model using CAD.	Ability to produce precise technical drawings using CAD; general literacy; basic technical/physical knowledge; ability to make autonomous design decisions based on a functional description and known parameters; spatial imagination.	Practical task (US) ²

Note: 1) PAL (2014_[24]); 2) Materials come from test blueprints made available by NOCTI at <https://www.nocti.org>; 3) <https://nictesting.org> 4) AKA (2021_[25]); 5) Zahnärztekammer Niedersachsen (2021_[26]); 6) PAL (2019_[27]).

The way forward

The project selected 13 tasks from six occupations for an exploratory assessment of AI and robotics performance on work tasks. The selected tasks cover diverse occupations and skills, allowing to test assessment methodologies in different set-ups. However, the selection represents only an excerpt of the wide occupational space. A comprehensive assessment of AI and robotics performance in occupations would require a much larger effort. The project will need to assess AI and robotics on a larger set of tasks that represent the variety of skills and knowledge used in the workplace. It should also attempt to study how AI progress in these skills changes occupational tasks and occupations.

A first step towards a comprehensive AI assessment is the development of a systematic approach for sampling occupations and occupational tasks. This sampling approach should rely on clearly defined criteria for selecting occupations. It should aim at producing a feasibly sized sample of work tasks that sufficiently represents key dimensions of human performance at work (e.g. knowledge, skills, abilities, proficiency). In addition, it should adequately cover various work contexts, spanning the different occupational and industry domains, as well as tasks of varying complexity and generality.

As one sampling challenge, not all occupations are subject to entry examinations. As described above, occupation entry examinations are less common in elementary occupations, agriculture, managerial and sales jobs (Koumenta and Pagliero, 2017_[4]). This may result in underrepresentation of these occupations in the assessment. To address this problem, the project will seek alternatives to entry examinations for unregulated occupations. One way would be to consult experts on the most common tasks in these occupations. Another way is to obtain examples of common tasks from occupational taxonomies, such as the Occupational Network (O*NET) database of the US Department of Labor (National Center for O*NET Development, n.d._[28]).

As another shortcoming, certification and licensing examinations are too focused on assessing professional skills. Consequently, they often neglect other types of skills that are equally relevant at the workplace, such as general cognitive skills or social skills. A systematic review on the methods of competence assessment in German VET showed that 60% of assessment instruments used in examinations targeted occupation-specific competences. However, only 24% assessed general competencies, such as writing, reading and mathematics. Another 9% focused on social competences such as communication skills (Rüschoff, 2019_[2]). This highlights the need for a battery of different instruments to assess AI and robotics, including education tests (Chapter 3) and AI benchmark tests (Chapters 6-8), to capture a wider array of skills.

In its third phase, the project will aim at developing methods for studying how occupations may change in response to evolving AI and robotics. AI and robotics will not simply substitute workers in occupational tasks. In some tasks, machines will support workers by completing only parts of the work. This may result in additional tasks for the worker, such as monitoring and supervising the computer systems, thus changing skill requirements for the job. Moreover, the application of AI and robotics may completely change the task by reinventing its solutions or by altering the work environment to adapt it to the use of machines.

To study the implications of evolving AI and robotics for occupations, the project will develop a set of task descriptions to illustrate how different human tasks are likely to evolve as we begin to carry them out with AI support. The set of tasks would need to illustrate the full range of tasks that are carried out at work and in everyday life, including tasks with key cognitive, physical and social aspects. For each of these tasks, the project will develop feasible scenarios for the way the tasks will likely be carried out with AI support. The goal is to have a group of AI experts, job analysts and psychologists analyse each of the sampled tasks to determine which activities could be performed by an AI system and then propose ways for redesigning the current task to allow a human to work with the support of an AI system. This would make it possible to describe a transformed role for humans in each of the tasks. The analysis would be carried out for each of the sampled tasks, considering current AI performance levels for the different capabilities, as well as several scenarios for future performance levels that AI could plausibly achieve in the next 5-20 years.

The next chapter describes the exploratory assessment of AI and robotics capabilities using expert evaluations on the selected 13 occupational tasks. It explores ways to address the challenges related to the use of occupational examinations for rating to develop a robust methodological approach for assessing AI and robotics performance in occupations.

References

- AkA (2021), *Abschlussprüfung Sommer 2021 Kaufmann/-frau für Büromanagement*, AkA Aufgabenstelle für kaufmännische Abschluss- und Zwischenprüfungen, IHK Nürnberg für Mittelfranken. [25]
- Ali, R. et al. (2023), *Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations*, Cold Spring Harbor Laboratory, <https://doi.org/10.1101/2023.03.25.23287743>. [18]
- Antaki, F. et al. (2023), "Evaluating the Performance of ChatGPT in Ophthalmology", *Ophthalmology Science*, Vol. 3/4, p. 100324, <https://doi.org/10.1016/j.xops.2023.100324>. [19]
- Badura, J. (2015), *Handlungsorientierte Aufgaben für schriftliche Prüfungen in der kaufmännischen Berufsausbildung - Erstellung und Korrektur. Leitfaden für Aufgabenersteller/-innen und Korretor/-innen*, [Action-oriented tasks for written examinations in commercial vocational training. A guide for task developers and graders.], AkA Aufgabenstelle für kaufmännische Abschluss- und, Nürnberg. [9]
- BIBB (2022), *Verzeichnis der anerkannten Ausbildungsberufe 2022*, (Directory of recognized training occupations 2022), Bundesinstituts für Berufsbildung, <https://www.bibb.de/dienst/publikationen/de/17944> (accessed on 27 October 2023). [6]
- BIBB (2013), *Empfehlung des Hauptausschusses des Bundesinstituts für Berufsbildung (BIBB) zur Struktur und Gestaltung von Ausbildungsordnungen (HA 158)*, [Recommendations of the Committee of the Federal Institute for Vocational Education and Training for the Structure and Design and Training Regulation Frameworks], Bundesanzeiger Amtlicher Teil (BAnz AT 13.01.2014 S1), <http://www.bibb.de/de/32327.htm> (accessed on 27 October 2023). [13]
- Buckendahl, C. (2017), "Credentialing", in *Testing in the Professions*, Routledge, New York, <https://doi.org/10.4324/9781315751672-1>. [12]
- Cedefop (2020), *Vocational education and training in Germany: Short description*, Publications Office of the European Union, Luxembourg, <http://data.europa.eu/doi/10.2801/329932> (accessed on 27 October 2023). [7]
- Eulerich, M. et al. (2023), "Can Artificial Intelligence Pass Accounting Certification Exams? ChatGPT: CPA, CMA, CIA, and EA?", *SSRN Electronic Journal*, <https://doi.org/10.2139/ssrn.4452175>. [22]
- European Commission (2008), *NACE Rev. 2 – Statistical Classification of Economic Activities in the European Community*, Office for Official Publications of the European Communities, Luxembourg, <https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF> (accessed on 27 October 2023). [16]
- Hambrick, D. and R. Hoffman (2016), "Expertise: A second look", *IEEE Intelligent Systems*, Vol. 31/4, pp. 50-55, <https://doi.org/10.1109/mis.2016.69>. [14]
- ILO (2012), *International Standard Classification of Occupations: ISCO-08*, International Labour Organization, Geneva. [15]
- Johnston, J. et al. (2014), "Determining BACB examination content and standards", *Behavior Analysis in Practice*, Vol. 7/1, pp. 3-9, <https://doi.org/10.1007/s40617-014-0003-6>. [1]

- Katz, D. et al. (2023), “GPT-4 Passes the Bar Exam”, *SSRN Electronic Journal*, [23]
<https://doi.org/10.2139/ssrn.4389233>.
- Kleiner, M. and E. Vorotnikov (2017), “Analyzing occupational licensing among the states”, [5]
Journal of Regulatory Economics, Vol. 52/2, pp. 132-158, <https://doi.org/10.1007/s11149-017-9333-y>.
- Koumenta, M. and M. Pagliero (2017), “Measuring prevalence and labour market impacts of [4]
occupational regulation in the EU”, report for European Commission, Directorate-General for
Internal Market, Industry,
[https://ec.europa.eu/docsroom/documents/20362/attachments/1/translations/en/renditions/nat](https://ec.europa.eu/docsroom/documents/20362/attachments/1/translations/en/renditions/native)
ive (accessed on 27 October 2023).
- Lin, J. et al. (2023), “Comparison of GPT-3.5, GPT-4, and human user performance on a [20]
practice ophthalmology written examination”, *Eye*, <https://doi.org/10.1038/s41433-023-02564-2>.
- National Center for O*NET Development (n.d.), *O*NET 27.2 Database*, [28]
<https://www.onetcenter.org/database.html> (accessed on 24 February 2023).
- NCSL (2022), *The National Occupational Licensing Database*, (database), [10]
<https://www.ncsl.org/labor-and-employment/the-national-occupational-licensing-database>
(accessed on 30 August 2023).
- Noever, D. and M. Ciolino (2023), “Professional Certification Benchmark Dataset: The First 500 [17]
Jobs For Large Language Models”.
- Nori, H. et al. (2023), “Capabilities of GPT-4 on Medical Challenge Problems”. [21]
- OECD (2023), *Is Education Losing the Race with Technology?: AI’s Progress in Maths and [3]
Reading*, Educational Research and Innovation, OECD Publishing, Paris,
<https://doi.org/10.1787/73105f99-en>.
- OECD (2021), *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, Educational [8]
Research and Innovation, OECD Publishing, Paris, <https://doi.org/10.1787/5ee71f34-en>.
- PAL (2019), *Abschlussprüfung Teil 1 - Technische(r) Produktdesigner(in) - Produktgestaltung [27]
und -konstruktion, Prüfungsprodukt, Musterprüfung*, PAL Prüfungsaufgaben- und
Lehrmittelentwicklungsstelle, IHK Region Stuttgart.
- PAL (2014), *Fachkraft für Metalltechnik - Leitfaden für die Zwischenprüfung - Musterprüfung*, [24]
PAL Prüfungsaufgaben- und Lehrmittelentwicklungsstelle, IHK Region Stuttgart.
- Rüschhoff, B. (2019), *Methods of competence assessment in vocational education and training [2]
(VET) in Germany – A systematic review*, Bundesinstitut für Berufsbildung,
<https://www.bibb.de/dienst/publikationen/de/17861> (accessed on 27 October 2023).
- US Department of Labor (2021), *Careeronestop*, website, <https://www.careeronestop.org/> [11]
(accessed on 30 August 2023).
- Zahnärztekammer Niedersachsen (2021), *Infos für Auszubildende und Ausbilder - [26]
Musterprüfungen*, <https://zkn.de/praxis-team/zan-beruf-und-bildung/ausbildung-zfa/infosauszubildende-ausbilder.html> (accessed on 10 November 2021).

Notes

¹ <https://www.ihk-aka.de/>

² Prüfungsaufgaben- und Lehrmittelentwicklungsstelle (PAL) <https://www.ihk.de/stuttgart/pal>

³ Zentralstelle für Prüfungsaufgaben <https://www.ihk-zpa.de/>

⁴ <https://zfamedien.de/zfa/>

5

Assessing AI capabilities on occupational tests

Margarita Kalamova, OECD

This chapter evaluates the capabilities of artificial intelligence (AI) in complex occupational tasks typical of real-world job settings. Using tasks from certification and licensing performance tests, the study aims to provide a more tangible assessment base than abstract constructs such as literacy and numeracy. Despite the clarity they offer, occupational tasks, given their complexity, pose methodological challenges in gathering expert judgements on AI's proficiency. Two pilot studies, containing 13 tasks across six occupations, revealed AI's aptitude in basic reasoning and language processing and limitations in nuanced and physically intricate activities. Expert feedback highlighted ambiguities in task descriptions and the difficulties of comparing AI and human skills. This chapter outlines the methodology, findings and implications of these assessments.

The AI Future of Skills (AIFS) project has extended the rating of artificial intelligence (AI) capabilities to complex occupational tasks taken from tests used to certify workers in different occupations. These tests present practical tasks that are typical in these occupations. As discussed in Chapter 4, this poses a clear advantage for gathering expert assessments on AI and robotics. Unlike assessments based on abstract constructs, such as general intelligence, broad content abilities (e.g. verbal, spatial, numerical abilities) or narrower abilities (e.g. perceptual speed, psychomotor abilities), occupational task evaluations provide meaningful insights into real-world scenarios and practical occupational behaviours. This offers a pragmatic and focused means to assess AI and robotics capabilities in specific occupational contexts.

The inherent complexity of these tasks means they differ from the questions in education tests used in the assessments discussed in Chapter 3. Occupational tasks require varied capabilities, often involving physical tasks, 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.

The AIFS project carried out two exploratory studies on the use of performance tests of occupational tasks for assessing AI and robotics capabilities. The project selected 13 tasks from six occupations, which were presented in Chapter 4, for an exploratory assessment of AI and robotics performance on work tasks. The selected tasks represent some important elements of reasoning, language and sensory-motor capabilities, a diverse set of work contexts and different levels of complexity. This allows the project to test assessment methodologies in different set-ups. The two studies explored the use of two distinct online surveys, different response formats and different instructions for rating expected AI and robotics performance on the example occupational tasks.

The results of the exploratory studies showed that AI performs well in areas of basic language processing and reasoning, efficiently handling tasks like retrieving specific terminology and ensuring grammatical accuracy. However, challenges emerge when tasks demand depth and nuance, such as synthesising knowledge for product development or leading patient interactions. Complexities remain in physical dexterity, especially in intricate manual tasks and interaction with human body parts. Controlled environments amplify AI's capabilities, but unpredictable settings highlight its current limitations, underscoring the need for further advancements.

However, the results also revealed some methodological challenges in collecting expert judgements on AI capabilities with occupational tasks. The feedback from experts unveiled ambiguities in task descriptions, a lack of clarity regarding the assumptions and a need for more contextual information in the first study. The second study attempted to map AI capabilities against human job requirements, and while experts commended the initiative, they faced significant challenges in the rating process. A primary concern raised was the ambiguous categorisation of AI capabilities needed for tasks. Moreover, the measurement scale introduced in the survey further exacerbated the confusion. The survey's structure also muddled the comparison between AI and humans, making it challenging for experts to assess AI's proficiency in certain tasks.

This chapter will first describe the process of collecting expert judgement on performance tests of occupational tasks. It will then present and discuss the results of the two assessments. Finally, it will include some thoughts about the way forward.

Collecting expert judgement on performance tests of occupational tasks

The method for collecting expert judgement

Two different assessments within a spell of three months were carried out, each with a separate online survey. These were followed by a group discussion among computer scientists with the participation of

two industrial-organisational psychologists. Each time, the participants took a week to complete the survey. During this period, they could access, re-access and modify their answers via an individualised link. In total, there were nine performance tests, containing 13 tasks, to rate.

A three-hour online group discussion took place a week after each of the online assessments. In each meeting, experts received detailed feedback on how the group rated AI and robotics abilities to take the various tests. Experts discussed the results, focusing on the performance tasks, on which there was some disagreement in the evaluation of AI and robotics performance. In addition, the experts provided feedback on the evaluation approach and described any difficulties in understanding and rating the survey questions.

In July 2022, the first exploratory study asked 12 experts to rate AI's ability to carry out 13 occupational assignments. This 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 covered diverse capabilities (e.g. reasoning, language and sensory-motor capabilities), occupations and working contexts. The materials describing the tasks varied in length and detail, which the project used to explore how different conditions for rating affect the robustness of the results.

In September 2022, a follow-up evaluation of the same tasks tested a new framing of the rating exercise. Experts rated 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. They could thus 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.

Both exploratory studies followed a behavioural approach for collecting expert judgement. As described in Chapter 2, this approach relies on few experts who engage in in-depth discussions to arrive at a consensus judgement on a question. This aims to address questions in their complexity by considering different arguments and perspectives, and drawing on the best of these arguments to build a group judgement.

Developing the questionnaires

The first study contained 13 occupational tasks stemming from nine German and US performance tests for occupations.

For each occupational task, experts were first asked, "How confident are you that AI technology can carry out the task?". The response options ("0% – No, AI cannot do it"; 25%; 50% – "Maybe"; 75%; "100% – Yes, AI can do it"; and "Don't know") combined their confidence and rating of the capability of AI. Specifically, "0% – No, AI cannot do it" meant the expert was quite certain that AI cannot carry out the task, while 25% meant "AI probably cannot do it". In answering this question, the study asked experts to have the final product/result in mind of that particular task, i.e. the assessment input/materials could be transformed to make them more "user friendly" for AI to carry out the task. The question aimed to understand whether AI can achieve the same results as humans independently of steps taken to achieve the results.

A second question asked: "**Humans would typically execute a number of subtasks while carrying out the task. Which of the following subtasks do you think AI can carry out independently?**". This aimed to understand whether AI can take over certain work processes and complement humans at the workplace in that particular task.

Most experts provided detailed explanations of their responses to each of the two questions and each occupational task.

The survey gave experts detailed instructions that defined the parameters for evaluating the potential use of AI and robotics on the 13 occupational tasks. In making their judgement for each task, experts were asked to consider "current" computer techniques. These would be any available techniques addressed sufficiently in the literature whose capabilities and limitations can be roughly described. The intent was to

include techniques whose capabilities have been demonstrated in research settings without worrying whether those techniques have been applied in any significant way. Experts could consider techniques that might need “reasonable advance preparation” to perform a particular occupational task. This advance preparation was to be considered applied research to prepare an existing technique with known capabilities for a new domain.

The follow-up study attempted to address some methodological issues encountered in the first study. Experts had pointed out that certain task descriptions lacked detail about the working context and boundary conditions for the tasks, requiring them to make speculative assumptions in their ratings. One task, for example, asked test takers to create a 3D solid model using computer-aided design software, without providing any information about the reference part (is it a cup, a car, etc.?) or how the task is to be carried out. To inform their judgements, some experts searched the Internet for explanations of work contexts. In particular, they sought work related to material sciences and engineering in technical occupations, and cosmetic and nursing procedures in the personal care industry.

To improve the task descriptions, some experts suggested the project should work with subject domain specialists and job analysts. The project would consider such collaboration in its explorations of task redesign, which is different stream of work from assessing AI and robotics capabilities. Instead, for the second study of occupational tasks, the project provided complementary videos and a revised job analysis of each task.

In the initial study, experts appeared to converge in their assessments regarding the capability needs of different tasks and the present proficiency of AI and robotics in these areas, which prompted the organisation of the subsequent study.

While the first survey had asked experts about their confidence that AI can carry out each specific occupational task, including a list of sub-steps, the second study asked them to rate the performance of AI on each task with regard to several categories of underlying capabilities. The categories of capabilities, 18 altogether, were borrowed from (McKinsey Global Institute, 2017_[11]) (consult Annex Table 5.A.1 presenting the capability scales).

In making their judgement for each capability and each task, experts could choose between three performance levels defined for each capability. They could also rate the particular capability as not needed for AI and/or humans for carrying out the particular task. The OECD had selected the few capabilities (out of 18) considered most relevant for the execution of each particular occupational task. Finally, the experts could indicate any other essential capabilities for each occupational task which they considered missing from the list of capabilities pre-selected by the OECD.

The feedback was mixed, suggesting this approach might have felt forced or possibly that the scales the project used were not optimal. Experts acknowledged the project’s effort in outlining occupational tasks and capabilities but pinpointed challenges in capability ratings. The capability categories, derived from McKinsey’s framework, were deemed unclear and inconsistently structured. The measurement scales of the capabilities also faced scrutiny for their ambiguity, especially around the human-level benchmark. Concerns arose regarding the questionnaire’s design, especially its alignment between AI and human-centred questions.

Evaluation of AI and robotics capabilities on tasks and subtasks

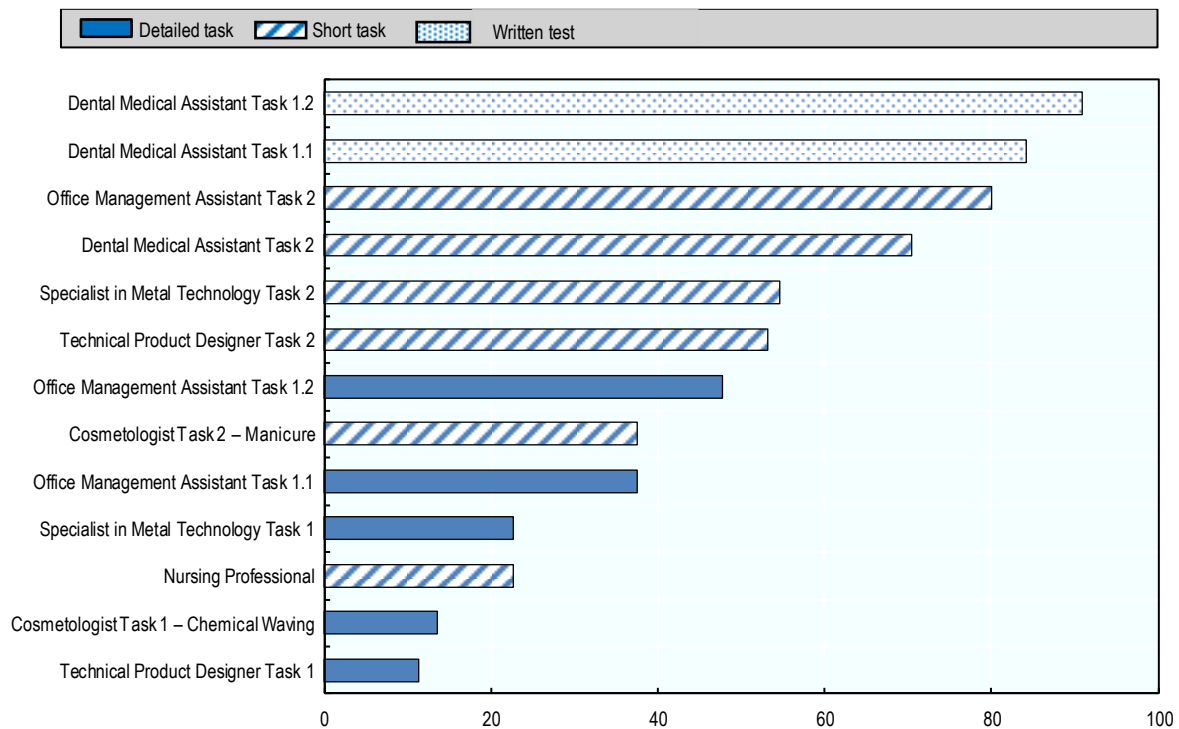
Average experts’ ratings of AI and robotics capabilities to carry out entire tasks

Figure 5.1 illustrates average measures of AI and robotics capabilities of carrying out the selected 13 occupational tasks. These measures are computed by taking the mean of the 12 experts’ responses to the question “**How confident are you that AI technology can carry out the task?**” for each of the 13 tasks.

“Don’t know” responses were excluded from these calculations. The measures thus show experts’ average confidence that a certain task can be entirely automated by AI and/or robotics systems.

Figure 5.1. AI and robotics performance on entire task, by task format

Mean of expert ratings to the question “How confident are you that AI technology can carry out the task?” (“0% – No, AI cannot do it”; “25%”; “50% – Maybe”; “75%”; “100% – Yes, AI can do it”; and “Don’t know”)



StatLink  <https://stat.link/1de8x3>

The average confidence measures vary significantly – from 10-92% – reflecting the diversity of represented occupations, their varying capability requirements and formats of exam tasks. Two of the tasks are written exam questions of a knowledge-based nature, while the rest are performance-based practical tasks to assess various ability domains. Knowing that AI systems have super-human performance on information retrieval tasks, it is not surprising that ratings for the Dental Medical Assistant tasks are notably higher than for tasks that require precise dexterity (Cosmetologist or Specialist in Metal Technology) and/or advanced reasoning (Technical Product Designer Task 1). Further down, the chapter will look more closely into the breakdown of the 13 tasks and plausible conditions and constraints for automation.

Some task descriptions are more complex and detailed than others, describing multiple sub-steps and providing instructions, which may also affect expert ratings. Most tasks with shorter descriptions in Figure 5.1 are rated higher than tasks with lengthy descriptions. As one possible explanation, shorter descriptions convey false simplicity because the brief explanation of the task may miss key points. This might be the case with the Technical Product Designer Task 2, which contains no detail about the type of final product and instructions on what needs to be done, making it appear simpler to carry out than the thoroughly described Technical Product Designer Task 1.

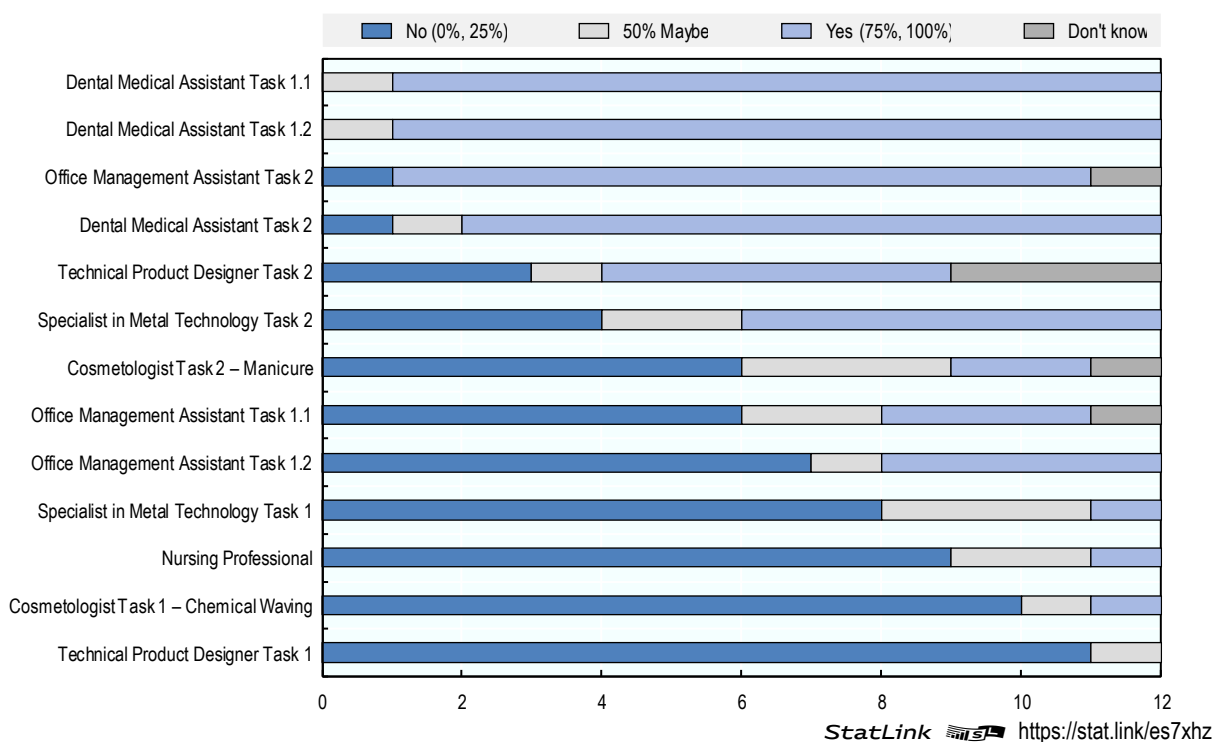
Another possibility is that shorter descriptions happen to refer to “simpler” tasks. For instance, when rating two similar tasks in the cosmetology occupation, experts had only 19% confidence, on average, that an AI or robotics system can carry out the thoroughly described task (chemical waving). They had 35%

confidence for the task with the short description (manicure). The higher rating for the task with a short description may reflect lower safety concerns and dexterity requirements, which indeed may make the task less demanding than the thoroughly described task involving a manipulation on a human head. It is difficult to draw conclusions about the potential bias in ratings arising from task descriptions. However, the project would need to carefully choose the right format and size of task descriptions for future assessments of occupational tasks.

Distribution of experts' ratings of AI and robotics capabilities to carry out entire tasks

Figure 5.2 provides important insights into experts' agreement on the various tasks. It shows the distribution of responses from 0% ("No, AI cannot do it") to 100% ("Yes, AI can do it"), whereas 0% and 25% are counted as No-answers and 75% and 100% as Yes-answers. The figure includes the "Don't know" answers as well. Following a simple majority rule – when seven or more experts provide the same No- or Yes-answer – a full consensus is reached on 9 of 13 tasks. Experts are confident about automating four tasks completely (those at the top of the figure). They are also confident that five other tasks (those at the bottom of the figure) are not fully feasible for AI and robotics systems yet. They disagree on the remaining four tasks: Technical Product Designer Task 2, Specialist in Metal Technology Task 2, Cosmetologist Task 2 and Office Management Assistant Task 1.1.

Figure 5.2. Distribution of expert ratings of AI and robotics performance on entire task



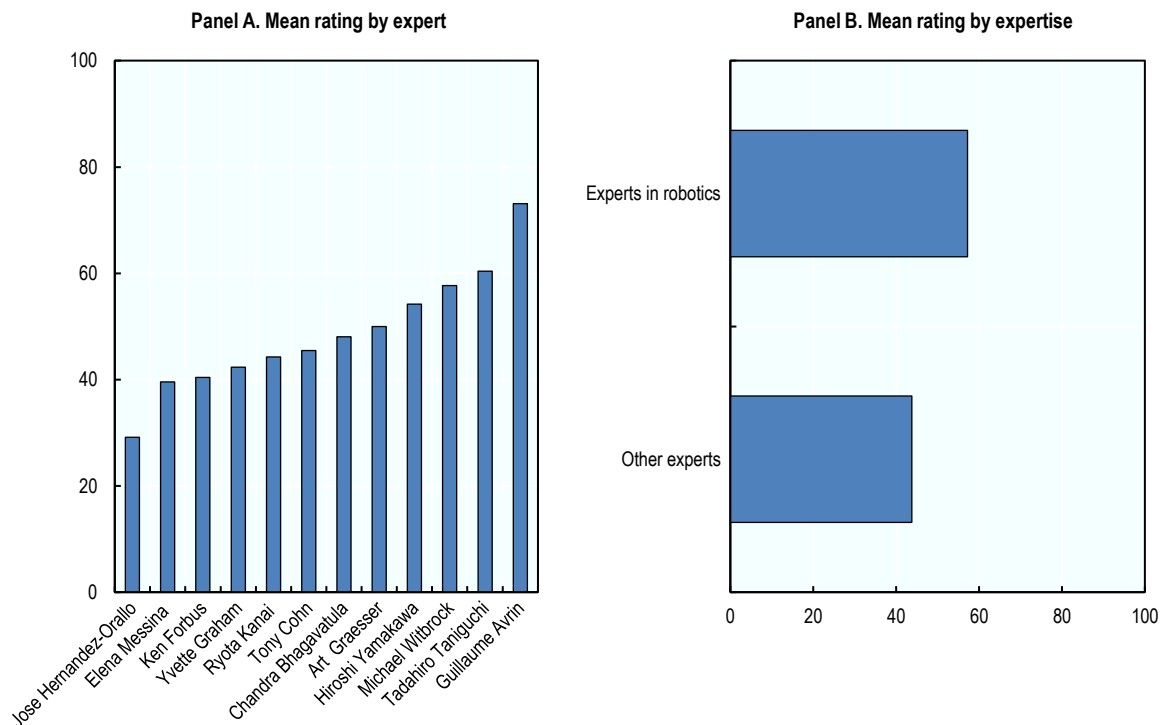
Ratings of AI and robotics capabilities to carry out entire tasks by expertise


Figure 5.3, Panel A shows the average ratings of each expert. These measures are computed by taking the mean of each expert's 13 responses to the question "How confident are you that AI technology can carry out the task?", one for each of the 13 tasks. "Don't know" responses were excluded from these calculations. The measures show experts' average confidence that AI and/or robotics systems can automate the selection of 13 diverse performance tasks. The results range from 30% for José

Hernández-Orallo up to 73% for Guillaume Avrin with the remaining ten experts having between 40-60% confidence about the bundle of 13 tasks.

Figure 5.3. Average AI and robotics performance, by expert and expertise

Mean of each expert ratings or expertise group ratings to the question “How confident are you that AI technology can carry out the task?” (“0% – No, AI cannot do it”; “25%”; “50% – Maybe”; “75%”; “100% – Yes, AI can do it”; and “Don’t know”) on all 13 tasks



StatLink  <https://stat.link/lyuvs7>

The 12 computer scientists come from different subfields of AI and robotics research. Four of the 12 experts – Guillaume Avrin, Tony Cohn, Elena Messina and Tadahiro Taniguchi – can be considered experts in robotics, while the remaining eight have a stronger expertise in disembodied AI. Although they all will most likely share the same knowledge on well-established techniques, each group may have specific expertise when it comes to new or less prominent approaches.

Figure 5.3, Panel B shows that the four robotics experts appear on average more confident about AI and robotics systems carrying out the bundle of 13 tasks than the other experts. However, due to the small number of observations (four robotics and eight other experts), these results need to be treated with caution. They do not necessarily mean that the robotics expertise is the driving factor. They may simply reflect differences across the whole group of experts, where a random selection of robotics experts happens to be rating the tasks more highly.

To further understand if robotics expertise was genuinely influencing the ratings, the project analysed average scores for various subtasks. These subtasks were divided into two broad categories: reasoning and language versus physical tasks that required dexterity, like those in robotic systems. To calculate the average scores for each subtask the project counted the number of “Yes”-responses to the question **“Humans would typically execute a number of subtasks while carrying out the task. Which of the following subtasks do you think AI can carry out independently?”** and then divided it by the total

number of experts (12). Subsequently, the project calculated two simple means per task: one on the subtasks in the physical domain and another on the subtasks in reasoning and language, the results of which are presented in Figure 5.4.

Figure 5.4. AI and robotics performance in broad capability domains, by task and expertise

Expert ratings of the question “Which of the following subtasks do you think AI can carry out independently?” (Yes/No answers) averaged by two broad capability domains (Reasoning/Language and Physical/Dexterity)

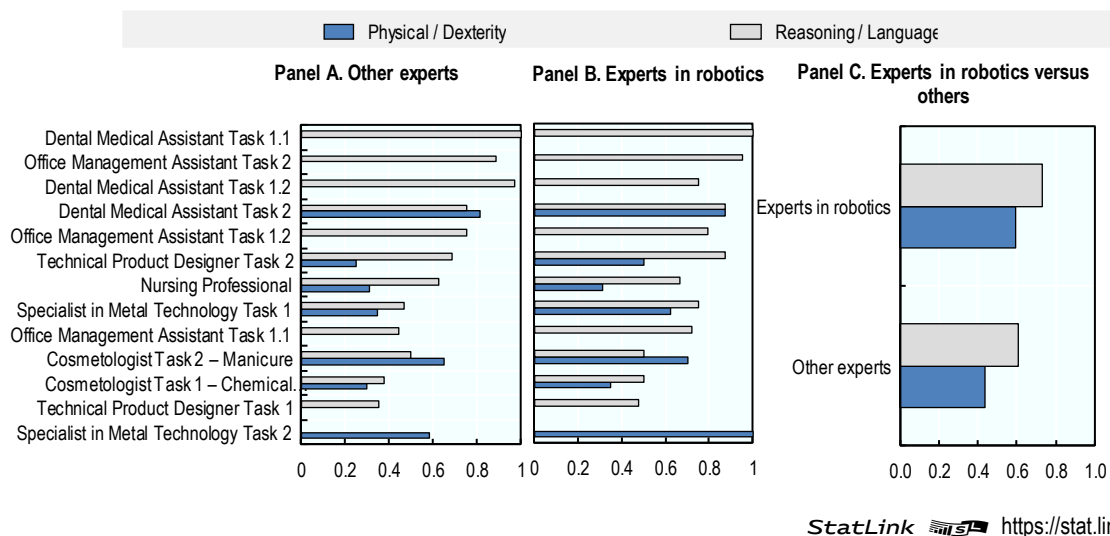


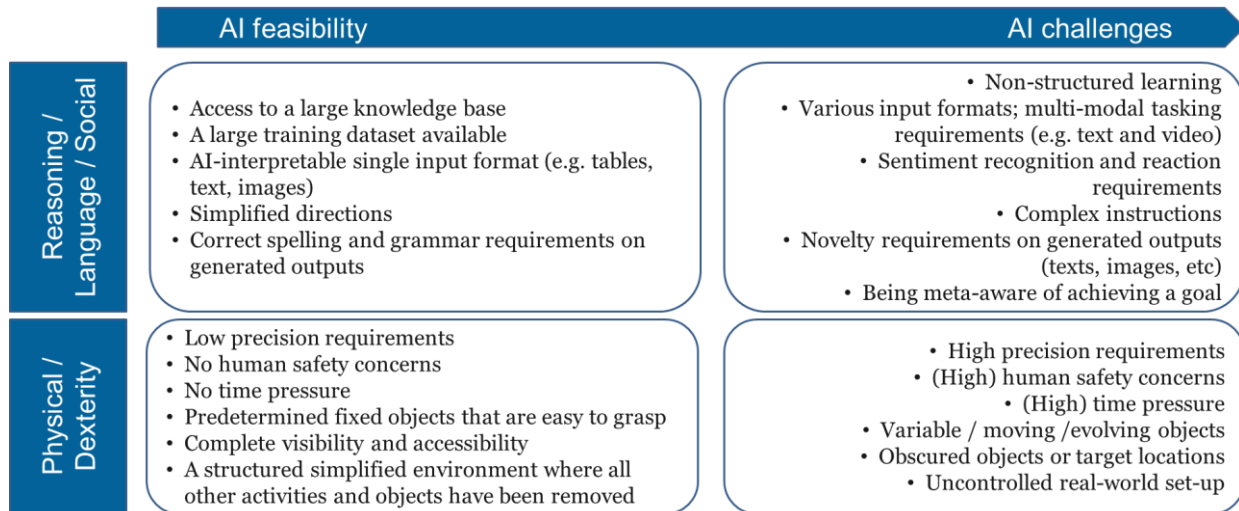
Figure 5.4 underscores the trend where robotics experts often rate tasks slightly higher, although there are exceptions. Physical tasks, especially those requiring dexterity, generally receive lower ratings compared to those centred around reasoning and language. This trend does not shift based on the presence or absence of robotics expertise, as shown in Panel C. However, when zooming into specific tasks, robotics experts exhibit confidence in AI's capacity to handle a large portion of the physical task within the areas of metal technology and product design. This might suggest that robotics experts are more optimistic in general. On the other hand, both roboticists and other experts display scepticism regarding AI's role in personal care tasks that involve comprehensive body movements, such as the Nursing professional role or the Cosmetology Task 1 (focusing on chemical waving). An exception here is the physical aspect of the cosmetology manicure task, which both groups believe AI can feasibly handle. The underlying reasons for these evaluations might revolve around safety considerations, the nature and quality of the target objects and other characteristics of the working environment.

What can and cannot AI systems do and under what conditions

AI's capabilities range from basic implementation to facing significant challenges, as demonstrated in Figure 5.5. By exploring distinct subtasks within the broad capability domains of reasoning and language and physical skills, a clearer picture emerges of where AI excels, where it performs moderately and where hurdles remain.

Figure 5.5. AI and robotics performance on subtasks, by complexity level and broad capability domain, mid-2022

Mean expert ratings of the question “Which of the following subtasks do you think AI can carry out independently?” (Yes/No answer). The subtasks listed in the boxes on the left have been rated as feasible by most of the 12 experts, while those on the right (AI challenges) have been rated as feasible by fewer than 5 experts.



In the domain of language and reasoning, AI represents varying degrees of proficiency. The computer scientists judged that AI could carry out basic subtasks, such as retrieving specific terminology. Examples include the task of a Dental Medical Assistant or using correct grammar and spelling in office documents. At the same time, more nuanced tasks such as “writing concisely and appropriately to addressee and purpose” and “documenting work steps” were judged as moderate challenges. Experts noted the greatest hurdles arise when AI was tasked with complex assignments such as synthesising knowledge into product design, communicating with patients or presenting novel ideas. This shows that while AI can process language, the depth and nuance of human reasoning remain a frontier.

The study took place in 2022 before the launch of ChatGPT, which appears to have meaningfully increased some AI language and reasoning capabilities. As highlighted by experts in follow-up meetings within the project, ChatGPT mimics language processing with more fluency and more contextual sensitivity than previous AI language systems. Moreover, its ability to simulate complex reasoning and human-like conversations signifies a marked improvement, bridging some of the subtasks experts initially identified as challenges in AI's mid-level mastery domain.¹

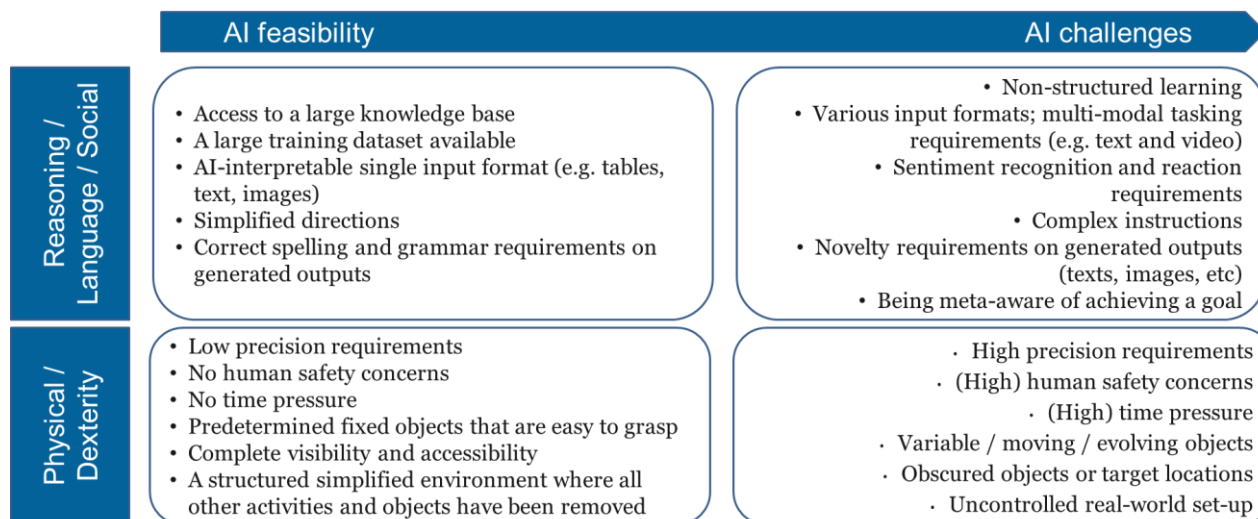
In the physical and dexterity domain, AI's performance varies based on task specificity and complexity. Basic procedural tasks, such as selecting the right materials or maintaining cleanliness in a cosmetology setting, are within AI's grasp. However, AI's mastery starts to waiver as tasks evolve in complexity, such as performing manicure techniques or assembling technical parts. The highest challenges are observed in tasks that require intricate manual skills and precision. This could include, for instance, correctly placing rods in human hair in cosmetology, or deburring and building a model of a part in metal machining.

Descriptors of capability levels of complexity

The scope of AI's capabilities is diverse, and a single subtask might be seen as easy or challenging, depending on the unique requirements and characteristics of a given workplace. Figure 5.6 presents and categorises certain requirements and characteristics that may either promote or deter automation, as outlined by experts. On the more feasible end, there are tasks where AI can easily be deployed, characterised by rule-based, structured environments. On the more challenging end, tasks that demand

meta-awareness, creativity or precise manual dexterity push the boundaries of current AI abilities. Between these extremes lies a spectrum of nuanced characteristics where AI can perform but also face obstacles and failures.

Figure 5.6. Expert descriptors of complexity levels of broad capability domains



A recurring theme in the expert discussion was the differentiation between tasks in controlled environments, like factories, and those in more arbitrary settings, like homes. Controlled environments allow for a higher degree of automation and predictability, making certain tasks seemingly more achievable for robots. In contrast, arbitrary environments present challenges in terms of variability, requiring higher levels of adaptability and dexterity from robots. Notably, there is a consensus that tasks like picking objects in cluttered spaces, often referred to as the "picking challenge", remain hard despite advances in robotics.

While robots can be highly specialised for particular tasks, their flexibility in handling variations or changes in tasks is still a challenge. The experts distinguished between highly specific tasks (like welding in a car manufacturing facility) and those that require a broader range of skills (like creating a work of art through welding). It is also crucial to know how much a system or environment needs to be engineered for a robot to successfully complete a task.

An essential factor was the role of robots in interacting with human body parts. Some experts lacked familiarity with cutting-edge robotics control technologies. However, they raised concerns about the complexities of ensuring safety when robots interact with the human body, particularly with current technology limitations. Furthermore, the current robotics technology still struggles with tasks involving dexterous manipulation, particularly when it comes to flexible materials like human hair. By contrast, if there is a low precision requirement and tasks involve fixed objects that are easy to grasp, AI could perform.

Time, often a luxury in professional domains, becomes an adversary for AI in tasks under time pressure. In metal technology, while AI can potentially handle welding or manufacturing, the need for swift, real-time decisions and actions can hamper its efficiency as noted by experts. As discussed above, the stakes rise when humans are the focus in medical emergency, necessitating AI to respond quickly and safely – a proficiency that remains underdeveloped.

Despite advances in large language models such as ChatGPT, certain challenges in language and reasoning identified above remain according to experts. In follow-up meetings, experts mentioned these models still grapple with non-structured learning and multi-modal tasks, such as processing varied input formats simultaneously; sentiment recognition can be hit-or-miss, and the models' ability to handle complex

instructions or ensure novelty in output is inconsistent. Moreover, experts lack consensus about whether these models truly possess meta-awareness regarding broader goals.

Discussion of the first assessment

The varied nature of the occupational tasks provided a broad view of different types of AI and robotics strengths and limitations, ranging from straightforward information retrieval to more intricate, multi-component activities.

Breakdown into subtasks

The experts greatly appreciated the task analysis and the breakdown of individual steps involved. This detailed segmentation of tasks into components provided clear insight into the challenges and requirements for AI and robotics. Many experts said this breakdown facilitated a more structured and nuanced understanding of the occupational tasks. While certain subtasks were deemed within the reach of AI, others remained elusive. This provided a more nuanced view on systems performance on a particular occupational task.

Experts raised the need for a more precise task analysis suited to AI and robotics. Some noted that human-centred job analysis might not suffice for AI evaluation. They suggested a detailed breakdown to focus on specifics that AI would need to emulate rather than generic human attributes like dexterity or strength. This suggests a deeper collaboration between job analysis experts and AI professionals.

Unclear assumptions

Given the high-level nature of the tasks, experts often formed their own assumptions, leading to potential inconsistencies in their ratings. They frequently highlighted the contrast between general-purpose and specialised AI systems. For some tasks, a general-purpose system, even with its robust capabilities, might find itself handicapped without specific prior data or training. Conversely, a specialised system might be more efficient but economically unviable due to high costs, especially when compared to human labour. As some experts insightfully noted, the nature of the task and its surrounding uncertainties determine the system's efficacy. For instance, a robot might seamlessly operate in a stable industrial environment. However, it might falter in more uncertain terrains, like personal services, without certain controls or constraints.

Further complexity arises when tasks demand multifaceted AI competencies. Some tasks, especially those necessitating fine dexterity, might require a combination of specialised AI algorithms for different components of the task. A system could entail a myriad of AI algorithms, each catering to specific facets like sensory processing, actuation and high-level task planning. In many instances, the hardware limitations of robots overshadow the cognitive capabilities of AI. Thus, separating these evaluations could lead to clearer insights.

The discussions underscored the importance of defining not only the nature of the AI system but also the environment within which it operates. Assumptions regarding environmental uncertainties can significantly impact the system's effectiveness. Explicitly clarifying these assumptions can streamline expert evaluations, ensuring they are premised on a shared understanding of the task, the AI system and the environment.

Complexity and pipeline architecture

There was a consensus that certain tasks presented in the rating exercise were highly complex, requiring the combination of multiple components or steps. Experts noted the challenge of chaining tasks together, especially in terms of error propagation. In a pipeline architecture, errors at one stage can compound,

leading to diminished overall performance. In tasks requiring multi-step object manipulation, for example, AI might handle individual steps efficiently. However, the accumulated uncertainty and error across multiple steps can compromise the outcome. Experts thought it might be valuable to explore and present tasks with alternative structures, such as parallel processing or hybrid models, to examine how AI and robotics perform under varied conditions.

Robotics considerations

When assessing tasks in the rating exercise, there is a notable distinction between the AI control mechanisms and the actual robotic capabilities. This distinction, though subtle, plays a pivotal role in the accurate evaluation of the feasibility and effectiveness of an AI-driven robotic system. The project did not provide any specific guidelines to experts on how to think about the level of robotics capabilities. Experts thus responded largely based on individual knowledge and understanding of contemporary robotics.

For many experts, the challenge arose not necessarily from the robotic capabilities side but more from a lack of clarity on the task requirements. As some experts were unfamiliar with tasks in areas such as metal technology and cosmetology, they admitted difficulty in matching up the task demands with existing robotic capabilities. This sentiment was shared even by those with expertise in robotics.

Another essential perspective brought forward was the importance of interpersonal interactions. For certain tasks, especially those in service sectors like cosmetology or nursing, technical performance is just one of several required dimensions. Experts highlighted interpersonal interaction as a critical part of these jobs. They underscored the need to consider the holistic requirements of an occupation and think beyond robotic capabilities.

Experts emphasised that while robotic systems capable of complex manipulations exist, they are not widely accessible. The difficulty of obtaining good robots for experimentation was noted as a significant barrier in many occupational contexts.

To provide a holistic picture, experts suggested that future exercises should consider including the current status of robotic hardware. Distinguishing between feasibility and limitations due to current hardware can be beneficial.

Lack of detail in some task descriptions

The feedback highlighted a desire for more context and detailed descriptions. Some experts felt the need to know more about the environment or specific task nuances. For instance, the “chemical waving” task did not consider hair types. Such information could significantly affect AI’s performance, as different hair types require varying product application times.

Experts noted that a more exhaustive breakdown of the tasks, considering various scenarios and nuances, might enable more precise ratings. Clarifying the specific environment, constraints and objectives would allow experts to rate capabilities based on a shared understanding in the future. Experts also suggested to enhance task descriptions with potential real-world variables. For instance, in tasks related to object manipulation, details about object weight, size and fragility can significantly influence the rating.

A significant feedback point was the need for visual aids or demonstrations to comprehend tasks better. For many, a brief video of an operator performing the task would provide a clearer perspective on the challenges and nuances. This idea extends to the suggestion that perhaps there could be an expert – a job analyst – on hand to answer questions or provide a brief overview.

Evaluation of AI and robotics capabilities on capability scales

As its primary aim, the second study explored diverse evaluation methods for the occupational tasks in response to feedback from the first survey. Experts rated potential AI performance with respect to several, pre-defined capabilities required for solving each task. The expectation was that linking occupational tasks to specific capability requirements would help experts abstract their evaluations from the concrete work context. In this way, they could focus more on general technological features needed for performing the task.

The study focused on the current state of AI technology and its ability to meet or surpass the complexities required of the tasks when carried out by humans. The first question of the new survey sought to measure the current capabilities of AI in relation to the task. In contrast, the second question aimed to understand the skillsets a human needs to perform the same task effectively. The underlying premise was to employ a scale for both AI and human capabilities, and then contrast these results. Ideally, an AI score above the human-required level on the scale would mean that AI could handle the task. However, the project recognised that AI might approach and solve the task differently, possibly without matching the exact complexity exhibited by humans. The survey, while looking at the capabilities of AI, also considered the potential for redesigning tasks, given the manner in which humans and AI tackle tasks may vary.

Of the initial 13 tasks, the project used only 9 for the second study; the other 4 were omitted due to irrelevance (written tasks) or inappropriateness for this more detailed exercise (tasks with limited descriptions).

Aggregate AI capability ratings

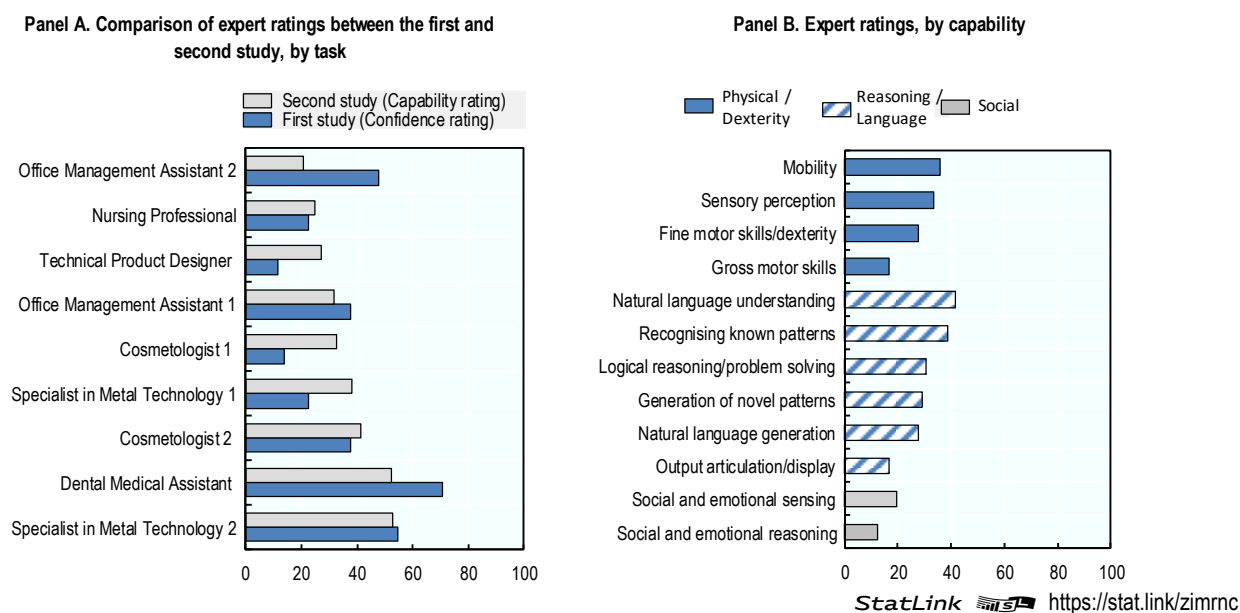
To evaluate AI capabilities based on the data from the second survey, the project determined three distinct aggregate indicators. The first measure considers all the essential capabilities for both entities for each task. It is calculated to represent the proportion of capabilities in which experts believe that AI's performance is equal to or surpasses the requirements set for human performance.

This evaluation was achieved by considering expert responses to two key questions in the survey. The first assessed the current capabilities of AI for specific occupational tasks ("In the context of this occupational task, what is the current AI capability in [particular capability]?"). The other determined the performance requirements for humans for those same tasks ("In the context of this occupational task and in your opinion, what are the requirements on humans in [particular capability]?"). The calculations excluded "Don't know" and 0 ("Capability not required for AI") responses.

The study then compared each expert's evaluation of AI capabilities with the corresponding evaluation of human requirements for each capability within each task. Whenever an expert judged AI's performance as superior, a score of 1 was assigned for that expert and capability within that task; otherwise, it was assigned 0. As a next step, the study calculates the percentage share of capabilities in a task that an expert considers equal or superior to the job requirements of the task. The aggregate measure is then constructed as the average of all experts' means for a particular task.

While this method assumes that all capabilities are equally important, it suggests that certain tasks might be achievable if most of the capabilities are met. However, this might not always be true. This method is a simplistic way of consolidating the evaluations. The resulting metric ranged between 0 and 100%, aligning it with the 0%-100% confidence scale from the first study of occupational tasks (Figure 5.7, Panel A). Obviously, the two measures are not fully aligned. Some tasks show the same characteristics, while others move in another direction.

Figure 5.7. AI capability expert ratings and their comparison to the ratings of the first study



The second metric represents the average confidence of experts that AI's performance is equal to or surpasses the requirements set for human performance in a particular capability domain across all tasks. It was calculated as the percentage share of expert scores of 1 (explained in the paragraph above) for each capability across all tasks. It thus aims to discern any logical distinction between lower-end and higher-end capabilities (Figure 5.7, Panel B).

Overall, experts are sceptical about AI performing at or superior to the job requirements in any of the broad capability domains. At the low end, there are factors like social and emotional sensing and reasoning. At the higher end, there are natural language understanding and recognition of known patterns in the reasoning/language domain and mobility in the physical abilities domain. However, the clarity of this perceived ordering at such a coarse level remains uncertain. In addition, the values for the four capabilities, Gross motor skills, Generation of novel patterns, Output articulation/display and Social and emotional reasoning, are based on ratings within the context of only one or two occupational tasks. Therefore, they should be considered with high caution.

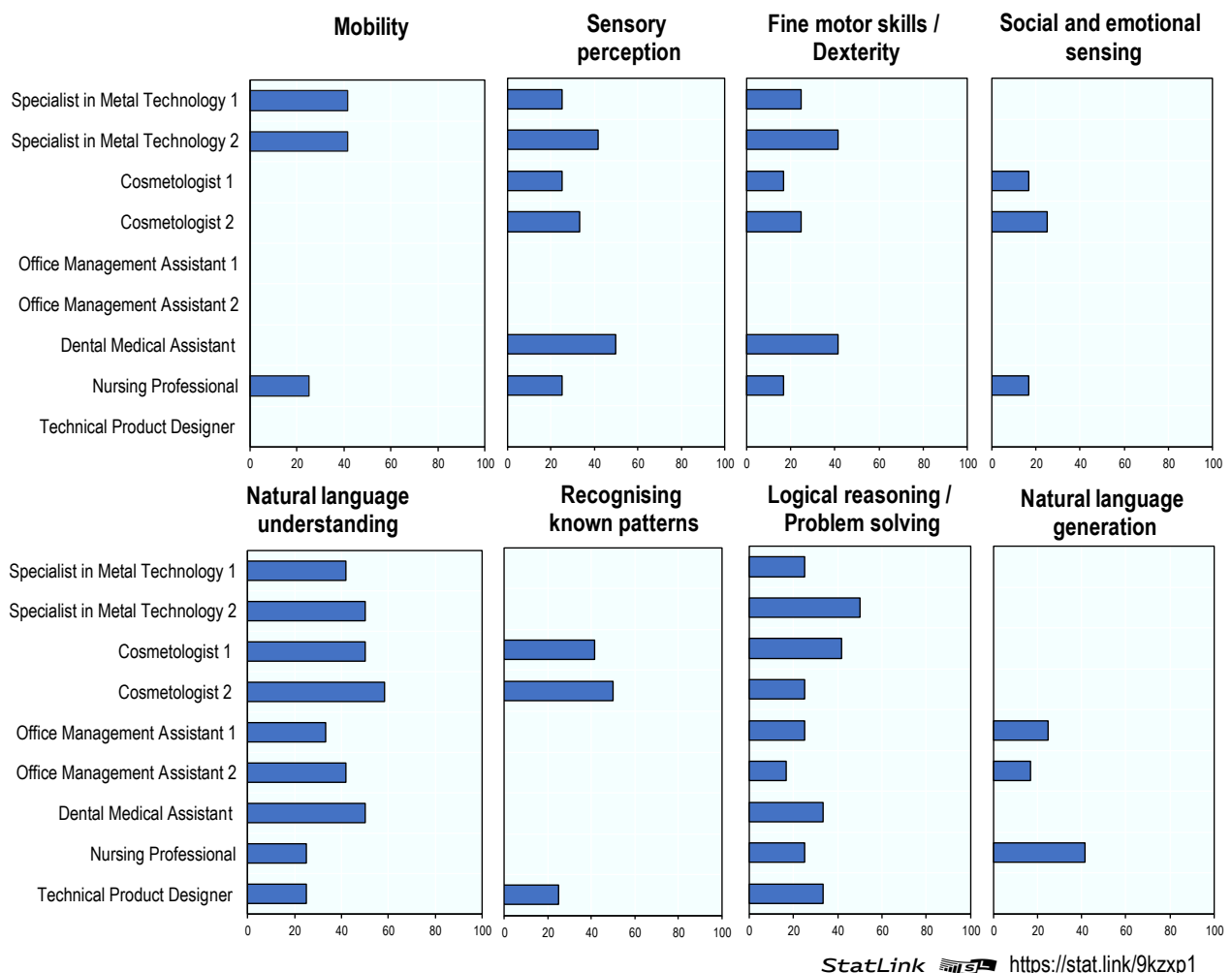
Figure 5.7 does not include the capabilities that experts identified as missing in their response to the third question of the online survey "Are there any essential capabilities missing from the above list?". Most experts generally did not dismiss the capabilities presented as being irrelevant to the task at hand.

AI capability ratings across task contexts

Figure 5.8 presents the analysis across various tasks using the study's third aggregate metric, which was calculated as the percentage share of expert scores of 1 for a particular capability and task. Interestingly, experts gave relatively different ratings for different capabilities across most of the tasks. This suggests the context of individual tasks had the expected impact on the ratings. Using the Nursing Professional Task as an example, there is a significant discrepancy between what AI can currently achieve and the requirements on human performance. Referring back to the initial survey's feedback, most experts believed AI was not adequately prepared to handle a large portion of the task. The comments particularly emphasised challenges related to the complexity inherent in NLP, the nuance of movements and the depth of sensory perception required. The Metal Technology Task 1 indicated similar challenges in the areas of sensory perception and dexterity as compared to the simpler Metal Technology Task 2. Meanwhile, the AI

capability ratings did not indicate a vast divergence for certain aspects such as natural language and mobility.

Figure 5.8. AI capability expert ratings, by task



Discussion of the second study

Experts appreciated the project's initiative to compare AI capabilities and human job requirements. They considered it a valuable starting point for understanding the potential and limitation of AI in occupational contexts. However, they faced many challenges in rating the capabilities.

Ambiguous categories of capabilities

Most experts considered the capability categories, as presented in the survey, vaguely defined and confusing. The definition of "natural language generation" illustrates their point well: it contains "web crawl result" as an anchor task, which is better aligned with information retrieval (a separate category) than natural language generation. Additionally, criteria such as whether the output should be realistic, logical or follow a command are missing. Meanwhile, terms like "nuanced language output" are ambiguous. Experts were concerned also with the organisation of categories. For instance, the distinction between navigation and mobility seemed superfluous, while it seemed surprising to group emotional and social sensing

together. More detailed definitions of the categories in the McKinsey framework might resolve these uncertainties.

The experts recommended the project develop an approach that uses clearer definitions for the different categories and provides tangible, real-world examples for each capability category. The categorisations and their subsequent groupings should be logical and intuitive to facilitate comprehension.

A confusing scale

The scale introduced in the survey was judged especially confusing. Experts noted the mention of quartiles did not correspond with the human-level benchmark, leading to uncertainty about what "human-level" genuinely meant. Moreover, the scale's metrics, including terms such as accuracy and complexity, were referenced without clear, defined thresholds, making it challenging to gauge the parameters. While certain categories did provide illustrative examples, like "picking up an egg", such examples were few. This scarcity left most categories without tangible references to calibrate the levels. Overall, experts found the labels arbitrary and problematic and felt confused about the proper use of the scales.

To foster greater clarity, experts highlighted the need to restructure the scales to include well-defined thresholds and distinctions between varying levels. They proposed introducing a nuanced scale, possibly leveraging a Likert statement. They also noted that most domains could likely benefit from at least five discernible capability levels, as the current state of AI often does not neatly fit into a single category. Especially when evaluating AI's dexterity, such as in manipulation skills – a domain where robots currently underperform – a more detailed scale becomes essential. Lastly, experts advised against using terms like "human" in labels and emphasised the importance of ensuring the scale's ends represent true opposites.

A questionnaire centred on AI versus humans

Some experts expressed concerns about the survey questionnaire not aligning its questions appropriately between those centred on machines and humans. The first question ("In the context of this occupational task, what is the current AI capability in [capability category] using the scale?") sought to determine AI's current capability. However, the subsequent question tried to identify the level deemed necessary for humans to carry out the task ("In the context of this occupational task and in your opinion, what are the requirements on humans in [capability category]?").

This differentiation raised concerns among some experts about the task assumptions. Did the survey envision a general-purpose humanoid robot designed to replicate human functions across various domains? Or did it envision specialised robotic systems tailored to specific tasks, such as adaptive devices to assist patients?

Furthermore, with respect to robotics operational independence, experts asked whether the robot would function autonomously after obtaining its occupational certification or serve as an assistant to humans. This was especially relevant in high-risk scenarios like heavy lifting or environments with extreme temperatures. Each perspective would change fundamentally the nature of the task.

Some experts also expressed confusion about whether they were rating an AI's overall ability in a specific capability domain or its competence in the context of a particular task. This dilemma was exacerbated when the task in question was relatively simple for humans but potentially complex for AI. The blending of these two perspectives in the instructions further complicated matters.

Experts agreed that merely determining if AI performs "better" than humans is not enough; they need to define what "better" means in terms of accuracy, speed or another metric. A main challenge they faced was comparing AI and human performance. They had to make assumptions during rating, which introduced variability in the responses. They expressed a strong need for clear guidelines when making comparisons, as different interpretations can significantly alter the results.

In turn, some experts disagreed with the response options to the first question, notably the “Capability not required for AI” option. They considered “capability not available in AI” as a more suitable response along the other three levels of AI capability (low, medium and high with descriptions). They also recommended to address the dynamics between AI and humans, ensuring the exercise captures the nuances of expectations placed on both sides. As a result of these ambiguities, many experts provided the “Don’t know” option to the first question.

Other experts noted that when social components were involved in the task, they raised their requirements on human performance. While the performance standard for simply completing a task might be similar for both AI and humans, expectations diverge when considering potential users or customers. They noted a general acceptance of certain limitations when it is known that AI performs a task, especially in social interactions or understanding. In contrast, for humans, the anticipation is considerably higher.

Useful videos

Experts found the videos accompanying the survey useful in understanding task complexities, particularly in areas unfamiliar to them. While these visuals conveyed the nuances and dexterity inherent in certain tasks effectively, the translation from instruction to action occasionally remained unclear. The videos underscored the challenges AI might face, yet some experts felt they mainly reinforced existing knowledge. While not deemed essential, the visual aids emphasised the intricacies of human roles and highlighted the challenges in adapting tasks for AI.

The feedback from experts has been mixed, indicating the new approach did not feel intuitive but also that the capability categories and scales the project used were not optimal.

The way forward

The two exploratory studies 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, judgements are required for all the required capabilities individually, as well as their combination. As a result, the project has decided to explore alternative uses of the occupational tasks.

The exercise provided important insights about how to think of and define the capability domains and suggested developing anchor tasks to describe each level of capability. The project will consequently explore working with the O*NET system of occupational classification. This provides specific tasks as anchors to help understand better each level’s capabilities as an alternative to the framework in this study that experts considered very general.

O*NET’s anchors serve as illustrative examples. This will make it easier for computer scientists and job analysts to agree upon the appropriate level for each task on the AI and human side, respectively. The O*NET system could provide clearer distinctions, especially in areas like natural language and fine motor skills. By presenting specific tasks for each capability level, it may be more intuitive and easier to comprehend than the broader categories in the current scale.

During the exercise, experts delved into the question of how AI can change the work context and suggested the use of occupational tasks to better anticipate how certain roles within the economy might evolve as new capabilities emerge. Experts highlighted the merit of exploring a human-AI collaborative approach where AI complements, rather than replaces (via automation) human efforts. Understanding these dynamics would be crucial for the goals of the project, ensuring that education, training and policy evolve hand-in-hand with technological advancements.

Experts provided useful advice on how to analyse task redesign. The project will draw on this advice in exploring the implications of evolving AI capabilities on education, work and everyday life. This exploratory work will consider the following points:

Rather than exclusively focusing on the current makeup of tasks, the design could contemplate the broader ecosystem within which these tasks exist. It is crucial to reflect upon how tasks can be reconceived, or entire systems revamped, to harness AI's strengths most effectively. Experts noted that while humanoid robots have allure, particularly from a human-computer interaction perspective, their development might not always be the most pragmatic or cost-efficient solution. In many scenarios, conceptualising the task or the system from scratch, with automation as a cornerstone, could yield higher efficiencies and superior user experiences. Expert reflections highlighted that such redesign decisions would be propelled by factors such as economic gains, consumer inclinations and technological breakthroughs.

Another significant observation stemmed from the potential disconnect between AI's capabilities and the specificities of the domain to which it is applied. While understanding the AI's capabilities is integral, having domain-specific knowledge is equally pivotal. To bridge this gap, some experts proposed a dyad approach. They felt a collaboration between an AI expert and a domain specialist could ensure a more holistic redesign of tasks that considered both AI's strengths and the intricacies of the domain.

References

McKinsey Global Institute (2017), *A Future that Works: Automation, Employment, and Productivity*.

[1]

Annex 5.A. Categories of AI capabilities

Annex Table 5.A.1. Categories of AI capabilities

Each capability category is characterised by three performance levels ranging from 1 (basic) to 3 (human-like) performance (based on tech advancements and complexity)

AI capability	1	2	3	Metric to define continuum
Natural language understanding	Low language comprehension required (while still accurate with structure commands)	Moderate language comprehension (medium accuracy of nuanced conversation)	High language comprehension and accuracy, including nuanced human interaction and some quasi language	Accuracy of comprehension Complexity of language/context integration
Sensory perception	Autonomously infers simple external perception (e.g., object detection, light status, temperature) using sensory data	Autonomously infers more complex external perception using sensors (e.g., high resolution detail, videos) and simple integration using inference	High human-like perception (including ability to infer and integrate holistic external perception)	Accuracy of perception/complexity of scene Degree of integration across sensors
Social and emotional sensing	Basic social and emotional sensing (e.g., object detection, light status, temperature) using sensory data	Comprehensive social and emotional sensing (e.g., voice, facial and gesture recognition-based social and emotional sensing)	High human-like social and emotional sensing	Quality of comprehension
Recognising known patterns/category (supervised learning)	Recognition of basic known patterns/categories (e.g., lookup functions in data modelling)	Recognition of more complex known patterns/categories	High human-like recognition of known patterns	Complexity of pattern
Generation of novel patterns/categories	Simple/basic ability for pattern/category recognition	More advanced capacity for recognition of new patterns/categories and unsupervised learning	High human-like recognition of new patterns/categories, including development of novel hypotheses	Complexity of pattern
Logical reasoning/problem solving	Capable of problem solving based on contextual information in limited knowledge domains with simple combinations of inputs	Capable of problem solving in many contextual domains with moderately complex inputs.	Capable of extensive contextual reasoning and handling multiple complex, possibly conflicting, inputs	Complexity of context and inputs
Optimisation and planning	Simple optimisation (e.g., optimisation of linear constraints)	More complex optimisation (e.g., product mix to maximize profitability, with constraint on demand and supply)	High human-like optimisation based on judgement (e.g., staffing a working team based on team/individual goals)	Degree of optimization (single vs. multi variate)

AI capability	1	2	3	Metric to define continuum
Creativity	Some similarity to existing ideas/concepts	Low similarity to existing ideas/concepts	No similarity to existing ideas/concepts	Novelty/ originality and diversity of ideas
Information retrieval	Search across limited set of sources (e.g., ordering parts)	Search across multiple set of diverse sources (e.g., advising students)	Expansive search across comprehensive sources (e.g., writing research reports)	Scale (breadth, depth, and degree of integration) of sources Speed of retrieval
Coordination With multiple agents	Limited group Collaboration; low level of interaction	Regular group interaction requiring real-time collaboration	Complex group interaction requiring high human-like collaboration	Complexity of coordination (i.e., number of interactions per decision) Speed/frequency of coordination
Social and emotional reasoning	Basic social and emotional reasoning	More advanced social and emotional reasoning	High human-like social and emotional reasoning	Complexity of emotional inference
Output articulation/ display	Articulation of simple content (e.g., organising existing content)	Articulation of moderately complex content	High human-like articulation	Complexity of message delivered. Variability in medium of message delivered
Natural language generation	System output with Basic written NLG (e.g., web crawl results)	System output with advanced NLP (more complex structure)	Nuanced, high human-like language output	Complexity of message delivered. Note: includes use of quasi linguistics (idioms, common names, etc.) Accuracy of audience interpretation
Emotional and social output	Simple social and emotional discussions (e.g., conversations with no gestures)	Advanced social and emotional discussions (e.g., conversations with gestures)	Nuanced high human-like body language and emotional display	Complexity of emotional communication Accuracy of audience interpretation
Fine motor skills/dexterity	Ability to handle and manipulate common simple objects (e.g., large solid objects) using sensory data	Can handle and manipulate wide range of more complex and delicate objects (e.g., pickup egg)	High human dexterity and coordination	Precision, sensitivity, and dexterity of manipulation
Gross motor skills	Basic 10/20 motor skills	More advanced multi-dimensional motor skills	High human multi-dimensional motor skills	Range and degree of motion Speed and strength of motion
Navigation	Use pre-defined algorithm for mapping and navigation	Autonomous mapping and navigation in simple environment	Autonomous mapping and navigation in complex environment	Complexity of environment (while still maintaining accuracy)
Mobility	Mobility/locomotion in simple environment (e.g., limited obstacles/office space)	Mobility/locomotion in more complex terrain of human scale environment (e.g., climbing stairs)	High human mobility and locomotion	Speed (gross motor) of mobility Scale of mobility vs.30 Complexity of environment/terrain

Source: McKinsey Global Institute (2017^[1]), *A Future that Works: Automation, Employment, and Productivity*.

Notes

¹ In the literature, there is uncertainty about the degree of generalisation reflected in the underlying language models that drive these AI systems and what that implies for the level of independent reasoning that the systems can carry out. In the context of this larger debate, the occupational tasks addressed in this chapter provide a special case. They occur in work settings where workers have been intentionally trained to carry out certain types of reasoning. Therefore, it makes sense to consider comparing those workers with AI systems that have been similarly trained on the reasoning required in that work setting.

6

A framework for characterising evaluation instruments of AI performance

Anthony G Cohn, University of Leeds

José Hernández-Orallo, Universitat Politècnica de València

Edited by: Sam Mitchell and Stuart Elliot, OECD

This chapter presents and discusses an approach to categorising benchmarks, competitions and datasets, jointly referred to as evaluation instruments of artificial intelligence (AI) performance. It proposes a set of 18 facets to distinguish and evaluate existing and new evaluation instruments, rating a sample of 36 evaluation instruments according to these facets. With a rubric composed of these 18 facets, four raters evaluate the sample, illustrating how well facets help analyse aspects of AI appraised by each evaluation instrument. In this way, the chapter proposes a framework that the OECD and third parties (researchers, policy makers, students, etc.) can use to analyse existing and new evaluation instruments.

Several studies focus on numeric comparison and the evolution of performance for a range of evaluation instruments of artificial intelligence (AI) (Martínez-Plumed et al., 2021^[1]; Ott et al., 2022^[2]). However, these studies only track the evolution of the progress of AI systems themselves. As such, they do not provide insight into how evaluation instruments such as benchmarks, competitions, standards and tests are also evolving. Nor do they indicate whether the measures are meeting the demands of a more comprehensive evaluation beyond some simple metrics. In response to this gap in the AI evaluation field, this chapter proposes a methodology to characterise the AI evaluation landscape. It will also assess the extent to which evaluation instruments can be used to evaluate the capabilities of AI systems over time.

There are thousands of evaluation instruments across all areas of AI, which makes it challenging to characterise the landscape of AI evaluation. As AI techniques evolve, they are also increasingly complex and diverse. Because of this, it is hard to analyse this evaluation landscape in a meaningful way. As a first step to overcome these challenges, this chapter presents and discusses an approach to categorising AI evaluation instruments. This categorisation is performed with a set of 18 facets, which are proposed to distinguish and evaluate the characteristics of existing and emerging evaluation instruments.

This chapter codes a sample of 36 evaluation instruments to evaluate how well the facets work in general and to what extent they help map the landscape of evaluation instruments and distinguish their differences. An evaluation instrument classification based on these facets may inform the design of future evaluation instruments. It is not clear if a single universal evaluation instrument will ever be feasible, or even a battery for each domain (vision, reasoning, etc.). Certainly, that ideal has eluded the community so far. The chapter aims to help direct future efforts in the evaluation of AI systems rather than find facet values that are valid for all evaluation instruments.

The 36 evaluation instruments classified in this chapter are only a cross-section of the thousands across all fields of AI research. Beyond the insights extracted from the sample, this paper and the rubric developed for the facets should serve as a reference for third parties (e.g. other researchers) to analyse other existing and newly proposed evaluation instruments. The work demonstrates that a set of evaluation instruments can be coded according to the facets in a relatively reliable manner. The resulting values reveal some interesting patterns about the characteristics of evaluation instruments used in the field.

The rest of the chapter is organised as follows. The second section presents the proposed 18 facets and a rubric that explains how facet values should be chosen. Then the criteria for selecting the 36 evaluation instruments and the methodology the raters used to apply the rubric is presented. The next section discusses the level of disagreement between raters for each facet and evaluation instrument. The penultimate section analyses the ratings of the 36 evaluation instruments, and what they reveal about this group of evaluation instruments. Finally, findings and possible future work is discussed in the final section.

Characterising AI evaluation instruments

The project initially hoped to find and build on existing methods to characterise evaluation instruments, but at the start of the project it became apparent a methodology that could be applied consistently across the AI evaluation field did not yet exist. Therefore, the project has defined a novel framework for this task inspired by work outside of AI research that has developed more systematic coverage of evaluation methods: the new set of facets proposed to evaluate evaluation instruments are inspired by psychological testing. The terminology used in this chapter is based on common use in AI, but also incorporates terms and concepts from the Standards for Educational and Psychological Testing by the American Educational Research Association (AERA, APA, NCME, 2014^[3]).

The following list proposes 18 facets to characterise existing and future evaluation instruments for AI. Each facet is followed by the values according to which an evaluation instrument can be classified in brackets. Some values indicate “(specify)”, which means the rater must respond in free text for that value. The colour

blue indicates cases where a facet has a preferred value, *in general* (for some evaluation instruments, another value may be preferred). However, some facets do not have a preferred value and therefore all values are left in black.

The facets are grouped into three main categories following the three main groups given by the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014^[3]): Validity, Consistency and Fairness. These groups deal with *what AI performance is measured*, *how it is measured* and *what AI system is measured*, respectively.

Validity facets (Does it measure what it should?)

- **Capability** [TASK-PERFORMANCE (specify), CAPABILITY (specify)]: does the evaluation instrument just measure observed (aggregated) performance on a TASK (e.g. protein folding, credit scoring) or can the evaluation instrument also measure a CAPABILITY (e.g. object permanence, dealing with negation)?
- **Coverage** [BIASED (specify), REPRESENTATIVE]: does the evaluation instrument cover a BIASED or unbiased (REPRESENTATIVE) distribution of what is meant to be measured?
- **Purpose** [RESEARCH, CONFORMITY, OTHER (specify)]: is the benchmark meant to foster research or development, or to certify whether an AI system conforms with some level or standard?
- **Realism** [TOY, GAMIFIED, REALISTIC, REAL-LIFE¹]: to what extent is the evaluation instrument a toy problem or a complex gamified problem? Is it a realistic setting (e.g. a simulated scenario, a lab or testing facility) or is the evaluation itself happening in real life?
- **Reference** [ABSOLUTE, RELATIVE (specify)]: are results reported as an absolute metric (criterion-referenced) or are they reported as a relative (percentage) metric to a reference (norm referenced), e.g. human performance?
- **Specificity** [SPECIFIC, CONTAMINATED]: are the results precisely aligned with what is meant to be measured or contaminated by other skills or tasks?

Consistency facets (Does it measure it effectively and verifiably?)

- **Adjustability** [UNSTRUCTURED, ABLATABLE², ADAPTIVE]: is the analysis of results on the set of instances unstructured?; has the evaluation instrument identified a set of meta-features such as difficulty or dimension that could be used to analyse the results by these dimensions (ablatability)?; or are these meta-features used to adaptively or adversarially choose the instances to test more informatively (adaptive)?
- **Containedness** [FULLY-CONTAINED, PARTIAL-INTERFERENCE (specify), NOT-CONTAINED (specify)]: Once started, is the testing isolated from external factors or interference possibly affecting results (human participants, online data, weather, etc.)?; is there some partial interference not affecting the results significantly?; or is it dependent on external resources and conditions?
- **Judgeability** [MANUAL, AUTOMATED, MIXED]: is scoring manual (e.g. through human questionnaires or judges) or automated (e.g. correct answers or optimality function) or a mixture?
- **Reliability** [RELIABLE, NON-RELIABLE, N/A]: does the evaluation present sufficient repetitions, episode length or number of instances to give low variance for the same subject when applied again (test-retest reliability)? If the testing methodology or the common use of the evaluation instrument is not clear, then N/A may be the most appropriate facet value.
- **Reproducibility** [NON-REPRODUCIBLE, STOCHASTIC, EXACT]: is the evaluation non-reproducible, with results biased or spoiled if repeated?; does the evaluation instrument have stochastic components leading to different interactions?; or are the results completely reproducible,

i.e. can the same exact test (inputs, interaction, etc.) be generated again for another (or the same) competitor?

- **Variation** [FIXED, ALTERED, **PROCEDURAL**]: is the evaluation based on fixed datasets?; have the instances been altered by adding post-processing variations (noise, rotations, etc.)?; or have the instances been created (e.g. using procedural generation³)?

Fairness facets (Does it treat all test takers equally?)

- **Ambition** [SHORT, LONG]: when the evaluation instrument was created, was it aiming at the short term (improving on the state of the art) or long term (more ambitious goals)?
- **Antecedents** [CREATED, RETROFITTED (specify)]: is it devised purposely for AI or adapted from tests designed to test humans?
- **Autonomy** [AUTONOMOUS, COUPLED (specify), COMPONENT]: is it measuring an autonomous system, coupled with other systems (e.g. humans) or as an isolated component?
- **Objectivity** [LOOSE, CUSTOMISED, **FULLY-INDEPENDENT**]: is it loosely defined, customised to each participant or does the evaluation instrument have a predetermined independent specification?⁴
- **Partiality** [PARTIAL (specify), **IMPARTIAL**]: does the evaluation instrument favour particular technologies, conditions or cultures that should not have an influence on the result of the evaluation?⁵
- **Progression** [STATIC, DEVELOPMENTAL]: Is the score measuring a capability at one moment or is it evaluating the development of the capability of the system within the test?

Facets with preferred values reflect suggestions about directions for changing the characteristics of evaluation instruments to improve them. For example, researchers testing AI should prefer an evaluation instrument that is **RELIABLE (Reliability)** to an evaluation instrument that is **NON-RELIABLE**, all things being equal. Facets that do not have preferred values are useful for categorising evaluation instruments in terms of other characteristics that may be useful for a particular purpose. For example, if a researcher is using an evaluation instrument that measures **TASK-PERFORMANCE (Capability)**, then they cannot draw conclusions about that AI's capabilities based on its performance on that evaluation instrument alone. An evaluation instrument that measures **CAPABILITY**, however, could be used to draw such conclusions.

Some of the facets, including across groups, are also closely related, such as {**Variation, Adjustability, Coverage**} or {**Objectivity, Reproducibility**}. One would expect that an evaluation instrument with a **FULLY-INDEPENDENT** value for **Objectivity** is more likely to be rated as **EXACT** for **Reproducibility**, for example.

Finally, the variability of measurement is also an important concept when evaluating evaluation instruments. In other words, how many changes can be made to an evaluation instrument for each AI system evaluation before the different evaluation results are no longer comparable? The term *accommodation* is “used to denote changes with which the comparability of scores is retained, and the term *modification* is used to denote changes that affect the construct measured by the test” (AERA, APA, NCME, 2014^[3]). This is important for **Specificity, Variation, Objectivity** and **Containedness**, as it indicates whether accommodations of the same test could evaluate different AI systems and even humans in a comparable way.

Evaluation instrument selection and rating methodology

Evaluation instrument selection

Evaluation instruments that met the following criteria were considered for inclusion:

- *Potential interest to understand the future of AI skills*: an evaluation instrument might be considered interesting if high AI performance can be regarded as indicating a noteworthy change in the capabilities of AI in general. In other words, progress in this evaluation instrument requires significant enhancement of AI techniques beyond the specific requirements of the evaluation instrument.
- *Diversity in the kind of task*: the evaluation instrument sample should cover a variety of domains (vision, natural language, etc.), formats (competitions, datasets, etc.) and types of problems (supervised/unsupervised learning, planning, etc.).
- *Popularity*: how many teams have already used this evaluation instrument? How many published papers refer to it? More popular evaluation instruments were preferred in the selection. Citations to the original papers introducing the evaluation instrument, the number of results on websites such as paperswithcode.com, etc., can be used as proxies to evaluate popularity. The possibility of industry-related evaluation instruments being less popular than research-oriented evaluation instruments was also considered.
- *Currency*: evaluation instruments still in active use or recently introduced were preferred rather than those that have fallen out of use.

The source of the evaluation instruments was mostly repositories⁶ and surveys, institutions such as National Institute of Standards and Technology and Laboratoire National de Métrologie et d'Essais, and competitions at AI conferences. The study then considered possible gaps and overlaps in the sample's coverage of domains. At the time of selection, only a rough estimate of potential preferred categories was possible for each evaluation instrument. The evaluation instruments have been categorised into six AI domains; the total count is more than 36 as multiple evaluation instruments tested AIs on more than one domain. For example, the Bring-Me-A-Spoon evaluation instrument (Anderson, 2018_[4]) evaluates AI on language understanding and robotic performance. The complete list of 36 selected evaluation instruments with their descriptions are shown in Annex Table 6.A.1.

Table 6.1. Primary testing domain of sampled evaluation instruments

AI domain	Reasoning	Language	Robotics	Vision	Video games*	Social-emotional
Number of evaluation instruments	12	11	7	5	6	1

Note: *Given the wide diversity of inputs across different video games and that different tasks within the same video game can require different capabilities, evaluation instruments based on video games were categorised separately.

These evaluation instruments cover a good distribution of benchmarks, competitions and datasets, although some can be considered to be in two categories. The term “test” to refer to an evaluation instrument is less usual. About half of the 36 evaluation instruments require use of language in the inputs and/or outputs, while about half require some kind of perception (mostly computer vision). There is some overlap in these two groups. Only a few evaluation instruments are related to navigation and robotics, in virtual (e.g. video games) or physical environments. A small number are related to more abstract capabilities or problems related to planning or optimisation.

Table 6.2. Type of sampled evaluation instruments

Evaluation instrument type	Competition	Benchmark	Dataset
Number of evaluation instruments	20	12	10

Note: Some evaluation instruments are a combination of types.

Rating methodology

A protocol to refine the rubric and to cover as many evaluation instruments as possible with available resources, is explained below. This protocol can be adapted to other situations or incorporate ideas from consensus-based ratings or the Delphi method (Hsu and Sandford, 2007^[5]). First, Anthony Cohn and José Hernández-Orallo acted as co-ordinators for the rating process, choosing four raters – Julius Sechang Mboli, Yael Moros-Daval, Zhiliang Xiang and Lexin Zhou (Cohn et al., 2022^[6]).⁷ Raters were AI-related undergraduate and graduate students and were recruited through a selection process, including interviews. Once the raters were appointed, each rater was given some meta-information about each evaluation instrument (acronym, name, major sources, what it measures, etc.) and completed other general information about each evaluation instrument (see Annex Table 6.A.1). They were also asked some information about their own completion, such as time taken (in hours). In all, 36 evaluation instruments were evaluated in this manner.

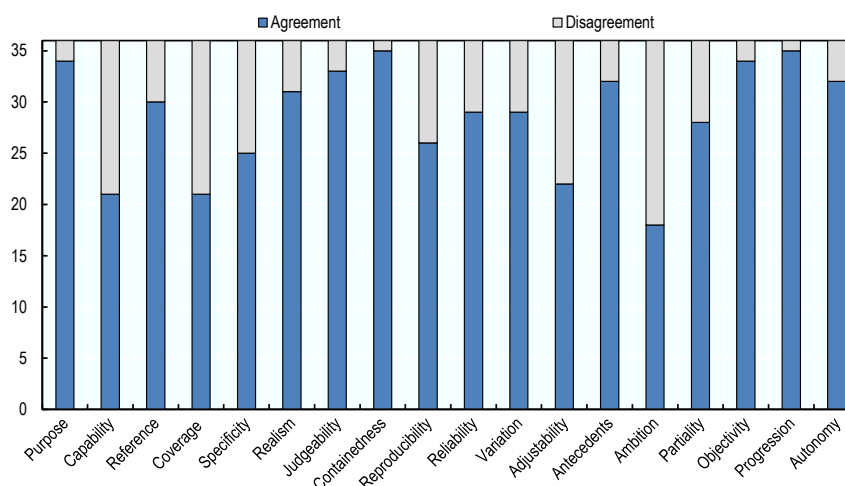
The evaluation instruments were rated in four batches. The first two evaluation instruments (Batch 1) were used by all co-ordinators and raters. All four raters then rated the next 11 evaluation instruments (Batch 2) and held discussions to refine the rubric. After observing consistent ratings across all four raters, only two raters rated each of the next ten evaluation instruments (Batch 3) and the final set (Batch 4) as this allowed a higher number of evaluation instrument evaluations with the same level of resources. Raters worked independently but discussed ratings after Batches 3 and 4, leading to some rating changes after the discussion. Annex Table 6.A.1 gives an overview of all 36 evaluation instruments and the batches they were evaluated in.

Analysis of rater consistency


The pattern of agreement or disagreement among the raters tends to vary depending on factors such as facet complexity, available information on the evaluation instrument and so on (see Figure 6.1). The most notable observations were the following:

- There is *consistent* agreement on **Progression, Autonomy, Purpose, Judgeability, Containedness, Objectivity** and **Autonomy** across all batches.
- There is *moderate* agreement on **Reference, Realism, Reproducibility, Variation** and **Partiality**. Notably, **Realism** has the largest number of values, but still obtains agreement well across evaluation instruments.
- There is the *least* agreement on **Capability, Coverage, Specificity, Adjustability** and **Ambition**, facets with mostly with binary options, with disagreement ranging from a third to a half of the evaluation instruments.

Figure 6.1. Rater agreement across all facets



Note: Agreements on facet value ratings for the 36 direct measures. “Agreement” means unanimous agreement and “Disagreement” covers all other cases.

StatLink  <https://stat.link/zk0q2p>

Overall, the results suggest that facets can be coded relatively reliably. Two factors help explain the lower rater consistency for some facets:

- To make justifiable decisions for facets like **Coverage** and **Specificity**, raters often needed to seek related literature for support when the answers were not clear from the specifications of evaluation instruments. Whether an evaluation instrument is specific (**Specificity**) and general (**Coverage**) enough for the measuring of certain capabilities is indeed hard to judge depending solely on the specifications. Furthermore, information extracted from different sources might lead to disagreements on selections.
- The subjectivity of a facet could also contribute to value divergences. This might be a reasonable explanation for inconsistent selections in **Capability**, **Adjustability** and **Ambition** since they allow raters more space for subjective interpretations. While relevant information regarding **Capability** and **Ambition** is often stated in the evaluation instrument specifications, these statements can somehow be interpreted in different degrees or ways. For example, an evaluation instrument for natural language understanding (NLU) could aim at improving state-of-the-art performance (short term) or measuring agents’ capabilities regarding NLU (long term); object recognition could be argued as a visual capability or a specific task.

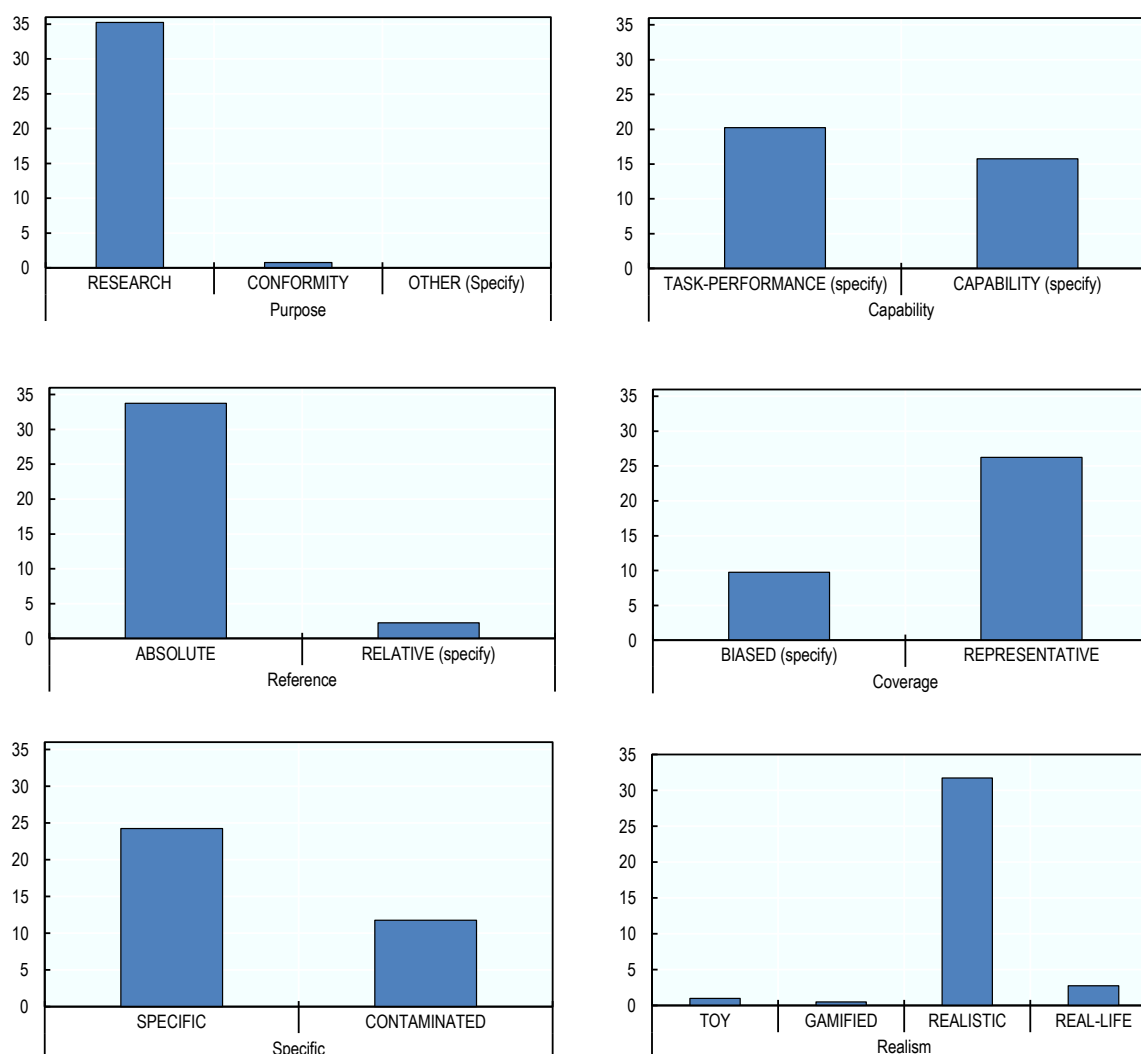
Analysis of facet values

Validity facets (Does it measure what it should measure?)

The findings indicate that sampled evaluation instruments are primarily designed for academic research, use absolute metrics and are divided roughly equally between measuring capabilities and specific performance tasks (see Figure 6.2).

- Nearly all the chosen evaluation instruments are aimed at promoting RESEARCH (**Purpose**) and predominantly use ABSOLUTE metrics (**Reference**).

Figure 6.2. Rater value selection on validity facets



StatLink  <https://stat.link/y0h5wp>

- The distribution of evaluation instruments measuring a specific task and those aiming for capability assessment is nearly balanced (**Capability**). This points to an ongoing debate in the field about the focal point of evaluation – performance or capabilities.
- Most evaluation instruments were classified as REPRESENTATIVE (**Coverage**). However, around 27% of the evaluation instruments are BIASED.
- About two-thirds of the evaluation instruments were SPECIFIC (**Specificity**). The remaining one-third were classified as CONTAMINATED, meaning the results may not fully align with the intended measurement objectives.
- Approximately 80% of evaluation instruments are REALISTIC (**Realism**), showing a strong inclination towards solving practical problems. Nonetheless, most evaluations have yet to be conducted in real-world settings.

The analysis found these areas for improvement:

- Most of the selected evaluation instruments that measure a capability (**Capability**) do not necessarily measure it reliably.

- Representativeness in the current evaluation instruments (**Coverage**) remains limited.
- Conducting more evaluations in real-world settings would promote development of more effective AI systems (**Realism**).

Consistency facets (Does it measure it effectively and verifiably?)

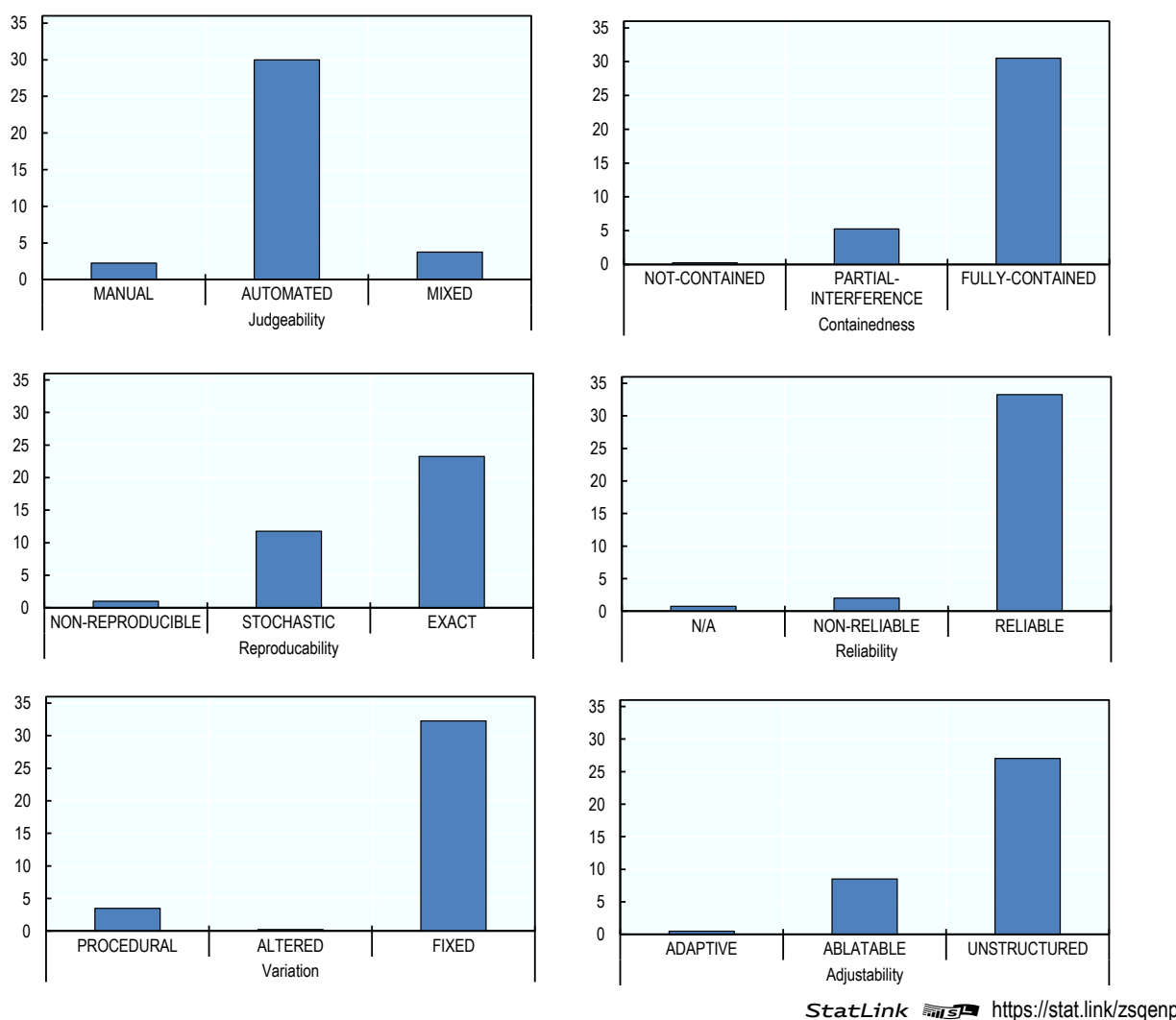
The results indicate that sampled evaluation instruments are mostly independent from external factors, are reliable, use fixed datasets and allow automated scoring (see Figure 6.3):

- Nearly all the selected evaluation instruments fall under the FULLY-CONTAINED category (**Containedness**), suggesting a high level of independence from external factors during assessments. This is a desirable feature for maintaining the integrity of an evaluation.
- Most evaluation instruments are classified as RELIABLE (**Reliability**), which lends credibility to the evaluation process.
- When it comes to **Judgeability**, most evaluation instruments employ AUTOMATED scoring instead of MANUAL or MIXED. While automated scoring generally offers more objectivity and speed, it does raise questions about the definition of the scoring metrics. For instance, determining the quality of a robotic dancer or cook through automated means can be challenging.
- In terms of **Variation**, nearly all evaluation instruments rely on FIXED datasets, which could limit the diversity in evaluation methods. For example, adding noise to the data could provide insights into the model's robustness.
- Most evaluation instruments are either UNSTRUCTURED or ABLATABLE (**Adjustability**), with few being ADAPTIVE. The absence of adaptive tests could be attributed to their operational complexity.

Further improvement is recommended in the following areas:

- introducing more diversity in the evaluation process, perhaps by adding post-processing variations or developing methods to cover intrinsic variations.
- encouraging more adaptive testing methods to evaluate how systems adapt to varying levels of difficulty.

Figure 6.3. Raters value selection on consistency facets



Fairness facets (Does it treat all test takers equally?)

The results show that most sampled evaluation instruments are impartial, objective and focus on static performance and AI systems working in isolation (see Figure 6.4):

- The raters found IMPARTIAL evaluation instruments account for 90% of the data (**Partiality**). However, the actual value might be lower since it is often hard to detect impartiality in the information given about an evaluation instrument. For instance, in an evaluation instrument for benchmarking clinical decision support systems, the training set may only include Latin American patients. However, the test set may include international patients.
- Virtually all the analysed evaluation instruments are classified as FULLY-INDEPENDENT (**Objectivity**), which favours fairness in evaluation.
- Nearly all evaluation instruments evaluate the AI systems are STATIC as opposed to DEVELOPMENTAL (**Progression**). This is possibly because many applications consider final performance as more important than how the system's performance evolves over time. It is also much harder to systematically evaluate AI systems over time. However, DEVELOPMENTAL evaluation instruments could give more insights into how the models learn with different data;

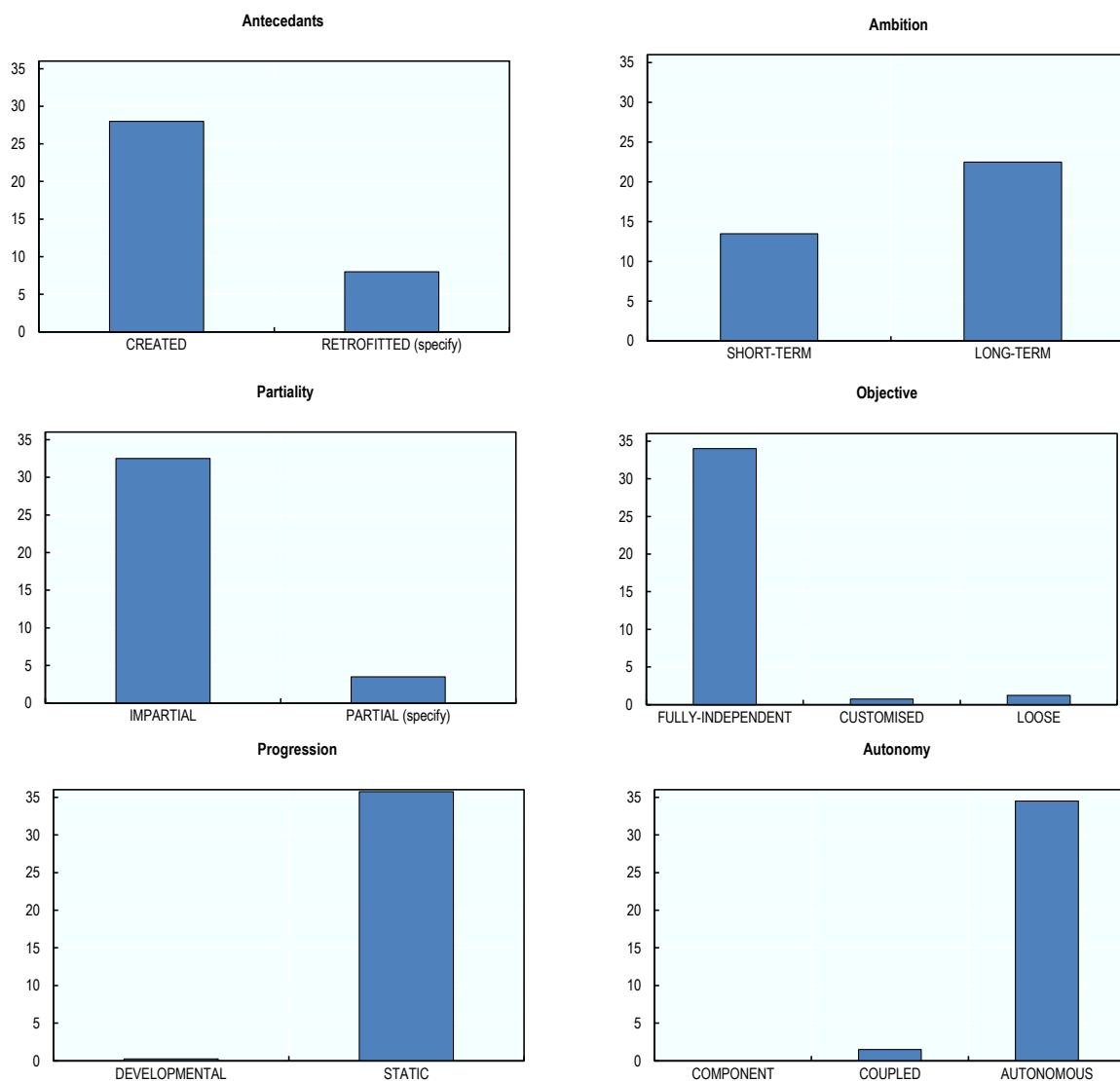
detect when and why things go wrong during the training phase; and identify the trade-off between training data size, time and performance.

- Only one of 36 evaluation instruments measured the performance of AI systems working with humans (COUPLED), rather than working autonomously or as an isolated component.

More efforts are needed to develop benchmarks that evaluate AI performance:

- at intervals through time as AI systems continue to develop (DEVELOPMENTAL).
- for those AIs that work together with humans (COUPLED).

Figure 6.4. Raters value selection on fairness facets



StatLink  <https://stat.link/e6jmcb>

Observations across facet groups

When looking at the distribution of facet values per evaluation instrument at a global level, evaluation instruments related to robotics and the physical world (Robocup, 2023^[7]; Robocup@Home, 2023^[8];

Lifelong Robotic Vision, 2023^[9]) have more variability in several respects. Many of them are judged manually as opposed to being automatically scored (**Judgeability**). Many have measures that are more realistic or close to real life (**Realism**). Testing is not always isolated from external factors but is often partially influenced by them (**Containedness**). In addition, they do not always measure systems autonomously, but sometimes with human interactions (**Autonomy**). One of the most popular evaluation instruments in the history of AI, ImageNet (Deng et al., 2010^[10]), is the only one which more than half of raters found to be partial (PARTIALITY), and the only evaluation instrument continuously rated as biased in **Coverage** (along with LibriSpeech). The disagreement in partiality may suggest that some sources of partiality are only discovered after the repeated use of an evaluation instrument and not identified by everyone immediately.

General Video Game Artificial Intelligence (Perez-Liebana et al., 2019^[11]) is a unique evaluation instrument that capitalises on the ablatable nature of video games, which can be altered easily by several characteristics or difficulty of the game. This is also going towards being procedural, but only to a limited extent as suggested by raters' values.

Finally, those evaluation instruments related to natural language, and especially WSC (Levesque, Davis and Morgenstern, 2011^[12]), GLUE (Wang et al., 2018^[13]), SUPERGLUE (Wang et al., 2019^[14]), Physical IQa (Bisk et al., 2020^[15]), SocialQA (Sap et al., 2019^[16]), SQUAD2.0 (Rajpurkar et al., 2016^[17]), WikiQA (Yang, Yih and Meek, 2015^[18]) and sW/AG (Zellers et al., 2018^[19]), have high degrees of contamination in the **Specific** facet. This might reflect the difficulty of isolating capabilities when using natural language, as some basic natural language competency requires many other things. This is reflected by the success of language models recently doing a variety of tasks (Devlin et al., 2018^[20]; Brown et al., 2020^[21]; Hendrycks et al., 2021^[22]; Bommasani et al., 2021^[23]), since mastering natural language seems to be contaminated by so many other capabilities and skills.

Conclusion

The framework presented in this chapter aims to provide a foundation from which evaluation instruments can be systematically evaluated and their evolution tracked.

The proposed set of facets and associated rubric, as well as the results of the study of 36 evaluation instruments reported in this paper, can be useful for three different kinds of users in slightly different ways.

1. First, evaluation instrument creators can see what design choices in their evaluation instrument to modify from a first evaluation of its facets and see how it compares to other evaluation instruments.
2. Second, AI system developers can choose the most appropriate evaluation instruments according to the facet values, and better understand what to expect from the evaluation and what it means exactly.
3. Finally, policy makers and stakeholders from academia, scientific publishing, industry, government and other strategic organisations can exploit an increasing number of evaluation instruments being evaluated and catalogued to understand the landscape of AI evaluation much better.

The facets framework can help these groups recognise gaps and limitations in evaluation instruments of AI performance. In this way, it helps stakeholders move beyond unstructured collections of benchmark results by metric, which are typical of the AI evaluation field. These can be useful for meta-analysis but are still lacking structure and insight about the evaluation instruments themselves.

The analysis of rater disagreement across facet values found it tended to reflect rater uncertainty about what evaluation instruments set out to measure or unresolved issues in AI evaluation. Section 4 observed disagreement between CAPABILITY and PERFORMANCE (**Capability**), between SPECIFIC and CONTAMINATED (**Specificity**), and between UNSTRUCTURED and ABLATABLE (**Adjustability**).

Evaluation instruments rated with the CAPABILITY (**Capability**) value were much more likely to be CONTAMINATED (**Specificity**). This may illustrate a difficulty in interpreting what the evaluation instrument designers intended to measure, particularly when measuring AI wider capabilities. The object of an evaluation instrument tended to be clearer to the raters when it was evaluating narrow task performance.

Rater disagreement may also be a sign of unresolved issues in AI evaluation: going from task-oriented evaluation based on performance to more general evaluation instruments lead to an evaluation instrument becoming CONTAMINATED (**Specificity**). For instance, adding many millions of examples can increase coverage. However, this adds problems of specificity and more difficulty in understanding the role each example plays in the overall score being measured by the evaluation instrument.

The most challenging parts of this proposal were:

- Determining the criteria for the inclusion of evaluation instruments.
- Defining facets that were difficult to understand or liable to be confused with others.
- Finding a protocol of application that is both sufficiently robust and can be used by a limited number of raters with restricted resources.

Finally, the categorisation framework for evaluation instruments presented here should be a living framework rather than set in stone. This would allow facets to be added, changed or removed and updated, and the rubric updated, to reflect the evolving nature of the evaluation of AI systems. However, some stability in names, facet values and facet description is needed to compile results of different rating studies over time. This would permit a large increase from the 36 evaluation instruments evaluated here to the order of hundreds in the future, with a more diverse and numerous pool of raters. Thus, rather than a continually evolving framework, it may be more sensible to review it periodically, following “change requests” from the community. A new, numbered version could be produced, with backward incompatibilities explicitly noted. Hopefully, these facets and the rubric describing them can help track the evolution of AI evaluation in the years to come, and identify the facets where changes are happening or should happen.

References

- AERA, APA, NCME (2014), *Standards for educational and psychological testing et al.*, American Education Research Association, https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf (accessed on 28 August 2023). [3]
- ANAC (2021), *ANAC2021 - 12th Automated Negotiating Agents Competition*, <http://web.tuat.ac.jp/~katfuji/ANAC2021/> (accessed on 18 October 2023). [41]
- Anderson, P. (2018), *Bring Me A Spoon | Matterport3D Simulator and Room-to-Room (R2R) data for Vision-and-Language Navigation*, <https://bringmeaspoon.org/> (accessed on 18 October 2023). [4]
- Assembly (2018), *Robotic Assembly – Recent Advancements and Opportunities for Challenging R&D | NIST*, <https://www.nist.gov/news-events/events/2018/08/robotic-assembly-recent-advancements-and-opportunities-challenging-rd> (accessed on 18 October 2023). [30]
- Bellemare, M. et al. (2013), “The Arcade Learning Environment: An evaluation platform for general agents”, *Journal of Artificial Intelligence Research*, Vol. 47, <https://doi.org/10.1613/jair.3912>. [24]
- Bisk, Y. et al. (2020), *PIQA: Reasoning about physical commonsense in natural language*, <https://doi.org/10.1609/aaai.v34i05.6239>. [15]
- Bommasani, R. et al. (2021), “On the Opportunities and Risks of Foundation Models”, <https://arxiv.org/abs/2108.07258v3> (accessed on 26 September 2023). [23]
- Brown, T. et al. (2020), “Language Models are Few-Shot Learners”, *Advances in Neural Information Processing Systems*, Vol. 2020-December, <https://arxiv.org/abs/2005.14165v4> (accessed on 26 September 2023). [21]
- Cohn, A. et al. (2022), “A Framework for Categorising AI Evaluation Instruments”, <http://ceur-ws.org/Vol-3169/> (accessed on 26 September 2023). [6]
- Deng, J. et al. (2010), “ImageNet: A large-scale hierarchical image database”, pp. 248-255, <https://doi.org/10.1109/CVPR.2009.5206848>. [10]
- Devlin, J. et al. (2018), “Bert: Pre-training of deep bidirectional transformers for language understanding”, *arXiv preprint arXiv:1810.04805*. [47]
- Devlin, J. et al. (2018), “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Vol. 1, pp. 4171-4186, <https://arxiv.org/abs/1810.04805v2> (accessed on 26 September 2023). [20]
- Froleyks, N. et al. (2021), “SAT Competition 2020”, *Artificial Intelligence*, Vol. 301, p. 103572, <https://doi.org/10.1016/J.ARTINT.2021.103572>. [28]

- Gaggl, S. et al. (2020), “Design and results of the Second International Competition on Computational Models of Argumentation”, *Artificial Intelligence*, Vol. 279, p. 103193, <https://doi.org/10.1016/J.ARTINT.2019.103193>. [26]
- Genesereth, M., N. Love and B. Pell (2005), “General Game Playing: Overview of the AAAI Competition”, *AI Magazine*, Vol. 26/2, pp. 62-62, <https://doi.org/10.1609/AIMAG.V26I2.1813>. [32]
- Harman, D. (1992), “Overview of the First Text REtrieval Conference (TREC-1).”, pp. 1-20, <http://trec.nist.gov/pubs/trec1/papers/01.txt> (accessed on 18 October 2023). [45]
- Hendrycks, D. et al. (2021), “Measuring Coding Challenge Competence With APPS”, <https://arxiv.org/abs/2105.09938v3> (accessed on 26 September 2023). [22]
- Hodaň, T. et al. (2018), “BOP: Benchmark for 6D object pose estimation”, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 11214 LNCS, pp. 19-35, https://doi.org/10.1007/978-3-030-01249-6_2. [42]
- Hsu, C. and B. Sandford (2007), “The Delphi Technique: Making Sense of Consensus”, *Practical Assessment, Research, and Evaluation*, Vol. 12, p. 10, <https://doi.org/10.7275/pdz9-th90>. [5]
- Levesque, H., E. Davis and L. Morgenstern (2011), “The Winograd Schema Challenge”, <http://www.aaai.org> (accessed on 26 September 2023). [12]
- Lifelong Robotic Vision (2023), *Lifelong Robotic Vision | IROS2019 Competition*, <https://lifelong-robotic-vision.github.io/> (accessed on 9 October 2023). [9]
- Linares López, C., S. Jiménez Celorrio and Á. García Olaya (2015), “The deterministic part of the seventh International Planning Competition”, *Artificial Intelligence*, Vol. 223, pp. 82-119, <https://doi.org/10.1016/J.ARTINT.2015.01.004>. [36]
- Maas, A. et al. (2011), *Learning Word Vectors for Sentiment Analysis*, <https://aclanthology.org/P11-1015> (accessed on 26 September 2023). [31]
- Marot, A. et al. (2021), “Learning to run a Power Network Challenge: a Retrospective Analysis”, *Proceedings of Machine Learning Research*, Vol. 133, pp. 112-132, <https://arxiv.org/abs/2103.03104v2> (accessed on 26 September 2023). [34]
- Martínez-Plumed, F. et al. (2021), “Research community dynamics behind popular AI benchmarks”, *Nature Machine Intelligence*, Vol. 3/7, pp. 581-589, <https://doi.org/10.1038/s42256-021-00339-6>. [1]
- MineRL (2023), *MineRL: Towards AI in Minecraft*, <https://minerl.io/> (accessed on 26 September 2023). [37]
- Mishra, S. et al. (2022), “NumGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks”, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 3505-3523, <https://doi.org/10.18653/v1/2022.acl-long.246>. [43]
- Ott, S. et al. (2022), “Mapping global dynamics of benchmark creation and saturation in artificial intelligence”, *Nature Communications*, Vol. 13/1, <https://doi.org/10.1038/s41467-022-34591-0>. [2]

- Panayotov, V. et al. (2015), “Librispeech: An ASR corpus based on public domain audio books”, [27]
ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Vol. 2015-August, pp. 5206-5210,
<https://doi.org/10.1109/ICASSP.2015.7178964>.
- PASCAL (2012), *PASCAL VOC 2012 test Benchmark (Semantic Segmentation) | Papers With Code*, <https://paperswithcode.com/sota/semantic-segmentation-on-pascal-voc-2012> [39]
 (accessed on 26 September 2023).
- Perez-Liebana, D. et al. (2019), “The Multi-Agent Reinforcement Learning in Malm\“O (MARL\“O) Competition”, <https://arxiv.org/abs/1901.08129v1> (accessed on 18 October 2023). [11]
- Rajpurkar, P., R. Jia and P. Liang (2018), “Know What You Don’t Know: Unanswerable Questions for SQuAD”, *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, Vol. 2, pp. 784-789, [33]
<https://doi.org/10.18653/v1/p18-2124>.
- Rajpurkar, P. et al. (2016), *SQuAD: 100,000+ Questions for Machine Comprehension of Text*, Association for Computational Linguistics, Stroudsburg, PA, USA, [17]
<https://doi.org/10.18653/v1/D16-1264>.
- Regnier, R. et al. (2021), “Validation de méthodologies d’évaluation de solutions de désherbage autonomes, dans le cadre des projets Challenge ROSE et METRICS.”, *Revue Ouverte d’Intelligence Artificielle*, Vol. 2/1, pp. 11-32, <https://doi.org/10.5802/ROIA.8/>. [38]
- Renz, J. et al. (2019), “AI meets Angry Birds”, *Nature Machine Intelligence 2019 1:7*, Vol. 1/7, [25]
 pp. 328-328, <https://doi.org/10.1038/s42256-019-0072-x>.
- RGMC (2022), *Robotic Grasping and Manipulation Competition @ ICRA 2022*, [46]
https://rpal.cse.usf.edu/rgmc_icra2022/ (accessed on 18 October 2023).
- Robocup (2023), *RoboCup Standard Platform League*, <https://spl.robocup.org/> (accessed on [7]
 26 September 2023).
- Robocup@Home (2023), *RoboCup@Home – Where the best domestic service robots test themselves*, <https://athome.robocup.org/> (accessed on 9 October 2023). [8]
- Sap, M. et al. (2019), “SocialQA: Commonsense Reasoning about Social Interactions”, *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 4463-4473, <https://doi.org/10.18653/v1/d19-1454>. [16]
- Sutcliffe, G. (2016), “The CADE ATP system competition - CASC”, *AI Magazine*, Vol. 37/2, [44]
 pp. 99-101, <https://doi.org/10.1609/AIMAG.V37I2.2620>.
- Vinyals, O. et al. (2017), “StarCraft II: A New Challenge for Reinforcement Learning”, [40]
<https://arxiv.org/abs/1708.04782v1> (accessed on 26 September 2023).
- Wang, A. et al. (2019), “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”, *Advances in Neural Information Processing Systems*, Vol. 32, [14]
<https://arxiv.org/abs/1905.00537v3> (accessed on 26 September 2023).

- Wang, A. et al. (2018), “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”, *EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop*, pp. 353-355, <https://doi.org/10.18653/v1/w18-5446>. [13]
- WN18RR (2023), *WN18RR Benchmark (Link Prediction) | Papers With Code*, <https://paperswithcode.com/sota/link-prediction-on-wn18rr> (accessed on 26 September 2023). [35]
- Yang, Y., W. Yih and C. Meek (2015), “WIKIQA: A challenge dataset for open-domain question answering”, *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pp. 2013-2018, <https://doi.org/10.18653/v1/D15-1237>. [18]
- Zellers, R. et al. (2018), “From Recognition to Cognition: Visual Commonsense Reasoning”, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2019-June, pp. 6713-6724, <https://doi.org/10.1109/CVPR.2019.00688>. [29]
- Zellers, R. et al. (2018), “SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference”, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 93-104, <https://doi.org/10.18653/v1/d18-1009>. [19]

Annex 6.A. Supplementary tables

Annex Table 6.A.1. Overview of Evaluation Instruments

Acronym	Reference	Type	Domain	Aim
WSC	(Levesque, Davis and Morgenstern, 2011 ^[12])	benchmark, competition	Reasoning	Targets evaluating common sense reasoning as a better alternative to the Turing test.
ALE	(Bellemare et al., 2013 ^[24])	benchmark	Video games	Intended to assess general AI using a variety of video games; exact metrics are unclear.
GLUE	(Wang et al., 2018 ^[13])	benchmark	Language	Measures AI performance in English natural language understanding tasks such as single-sentence tasks, similarity, paraphrasing and inference.
SUPERGLUE	(Wang et al., 2019 ^[14])	benchmark	Language	Measures AI performance in English natural language understanding tasks such as single-sentence tasks, similarity, paraphrasing and inference.
IMAGENET	(Deng et al., 2010 ^[10])	competition	Vision	Assesses AI's visual recognition abilities in object recognition, image classification and localisation amid varied conditions.
AIBIRDS	(Renz et al., 2019 ^[25])	competition	Video games	Evaluates an agent's planning ability in large action spaces by making it play Angry Birds.
ICCM	(Gaggl et al., 2020 ^[26])	competition	Reasoning	Compares the performance finding logical solutions in argumentation tasks.
Robocup	(Robocup, 2023 ^[7])	competition	Robotics	Aims to advance multi-robot systems through soccer matches.
Robocup@home	(Robocup@Home, 2023 ^[8])	competition	Robotics	Assesses AI robots in delivering assistive services for future domestic use.
Librispeech-SL12	(Panayotov et al., 2015 ^[27])	dataset	Language	Provides a free English speech corpus for training/testing speech recognition systems.
GVGAI	(Perez-Liebana et al., 2019 ^[11])	competition	Video games	Targets systems that can excel in multiple video games as a step towards artificial general intelligence.
PIQA	(Bisk et al., 2020 ^[15])	benchmark, dataset	Language, Reasoning	Evaluates language-based physical interaction reasoning for both typical and unconventional object uses.
SAT	(Froleyks et al., 2021 ^[28])	competition	Reasoning	Focuses on improving the performance and robustness of SAT solvers.
VCR	(Zellers et al., 2018 ^[29])	dataset	Reasoning, Vision	Identifies human actions and goals from visual cues.
Assembly	(Assembly, 2018 ^[30])	competition	Robotics	Assesses robotic systems' competencies using formal evaluations to guide development and match user needs.
IMDb	(Maas et al., 2011 ^[31])	dataset	Language	Detects text sentiment.
SocialQA	(Sap et al., 2019 ^[16])	benchmark	Socio-emotional	Measures computational models' social and emotional intelligence through multiple-choice questions.
GGP	(Genesereth, Love and Pell, 2005 ^[32])	competition	Video games	Tests AI's ability to play multiple games.
SQUAD2.0	(Rajpurkar, Jia and Liang, 2018 ^[33])	dataset	Language	Evaluates reading comprehension abilities.

Acronym	Reference	Type	Domain	Aim
ZellersWikiQA	(Yang, Yih and Meek, 2015 ^[18])	benchmark	Language	WIKIQA is a dataset for open-domain question-answering.
sW/AG	(Zellers et al., 2018 ^[19])	dataset, benchmark	Language, Reasoning	Measures grounded commonsense inference by answering multiple-choice questions.
L2RPN	(Marot et al., 2021 ^[34])	competition	Reasoning	Tests AI's ability to solve an important real-world problem for the future.
W/AG Lifelong-Robots	(Lifelong Robotic Vision, 2023 ^[9])	competition	Robotics, Vision	Tests AI's ability to solve an important real-world problem for the future.
WN18RR	(WN18RR, 2023 ^[35])	dataset	Reasoning	Measures success in link prediction tasks without inverse relation test leakage.
Planning	(Linares López, Jiménez Celorrio and García Olaya, 2015 ^[36])	competition	Reasoning	Assesses automated planning and scheduling across different problem families.
MineRL	(MineRL, 2023 ^[37])	competition	Video games	Evaluates the performance of reinforcement learning agents in playing Minecraft.
ROSE	(Regnier et al., 2021 ^[38])	competition	Robotics (non-humanoid)	Measures agricultural robotics' market-related aspects for the near future.
PASCAL-VOC	(PASCAL, 2012 ^[39])	dataset, competition	Vision	Evaluates computer vision tasks like object detection and image segmentation.
Starcraft II	(Vinyals et al., 2017 ^[40])	benchmark, dataset	Video games	Assesses agents in playing Starcraft II, also examines perception, memory and attention.
ANAC	(ANAC, 2021 ^[41])	competition	Language, Reasoning	Evaluates multi-issue negotiation strategies.
BOP	(Hodaň et al., 2018 ^[42])	benchmark	Vision	Measures "object pose estimation" in RGB-D images.
NumGlue	(Mishra et al., 2022 ^[43])	benchmark	Reasoning	Evaluates basic arithmetic understanding.
CASC	(Sutcliffe, 2016 ^[44])	competition	Reasoning	Evaluates theorem proving.
TREC	(Harman, 1992 ^[45])	competition	Language	Evaluates information retrieval technology through adaptive yearly competitions.
Bring-MeASpoon	(Anderson, 2018 ^[4])	Benchmark, dataset	Language, Robotics	Tests an agent's ability to navigate to a goal location in an unfamiliar building using natural language instructions.
RGMC	(RGMC, 2022 ^[46])	competition	Robotics	Assesses robotic grasping and manipulation capabilities.

Note: The yellow shaded evaluation instruments are Batches 1 and 2, green shaded items Batch 3 and blue shaded items Batch 4.

Notes

¹ REAL-LIFE does not mean a final or specific product in operation. It can also happen in early stages of research, such as evaluating prototype chatbots in a real social network.

² In AI research, the term “ablatable” refers to a component or feature of a system that can be removed or “ablated” to assess its impact on the system's overall performance.

³ Although PROCEDURAL was coloured, procedural may not always be better and can lead to problems if variations are not in an appropriate proportion. Also, generated data may just lead to a learning algorithm reverse-engineering the generator.

⁴ LOOSE refers to cases when evaluation is open, e.g. a robotic-domain evaluation instrument where a satisfactory interaction with the user is evaluated, but not even a clear questionnaire is defined. FULLY-INDEPENDENT could treat different groups differently if there is a reason for equality of treatment.

⁵ Coverage is about the domain, while Partiality is about how the evaluation instrument may favour some test-takers over others.

⁶ Repositories used were: Papers with code (<http://paperswithcode.com>), Kaggle (<http://kaggle.com>), Zenodo (<https://zenodo.org/record/4647824#.YV7CPdrMKUk>), Electric Frontier Foundation (<https://www.eff.org/ai/>), Wikipedia (https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research), Challenges in Machine Learning (<http://www.chalearn.org>).

7

AI direct tests: LNE and NIST evaluations

Elena Messina, Prospicience LLC, formerly: National Institute of Standards and Technology

Guillaume Avrin, Laboratoire National de Métrologie et d'Essais

Swen Ribeiro, Laboratoire National de Métrologie et d'Essais

Edited by: Abel Baret, OECD

Artificial intelligence (AI) has developed significantly in recent years. Its increased application in the industrial and domestic worlds raises questions about how it complements human intelligence. It seems only possible to evaluate this complementarity task by task or capability by capability. This chapter proposes a method and criteria (nature of the evaluation task, application area, level of difficulty, etc.) for systematising tasks on which AI and robotics systems have been evaluated in the past. This will allow the extraction of areas already covered and those yet to be evaluated. This method is applied to evaluation campaigns by the National Institute of Standards and Technology in the United States and the French Laboratoire National de Métrologie et d'Essais over the last decades. The paper concludes with a proposal for next steps to complete the mapping based on expert judgement.

Artificial intelligence (AI) has developed significantly in recent years, thanks especially to more advanced algorithms, easier access to data and greater computing power. The deployment of these intelligent technologies is under way in many spheres. In the professional world, for example, AI is used in inspection and maintenance robots, collaborative industrial robots and agricultural robots. In private life, AI manifests in technologies such as personal assistance robots, autonomous vehicles and intelligent medical devices.

As a result, public policies dedicated to AI technologies are emerging. These aim to facilitate AI development (Van Roy, 2020^[1]) and authorise their deployment (see European Commission (2021^[2])). They also aim to ensure their sustainability, such as in the US plan for technical standards and tools (NIST, 2019^[3]). A thorough understanding of AI capabilities and their link to human skills is needed to guide design of such policies.

The AI and Future of Skills (AIFS) project is exploring different ways and methodologies to develop comprehensive measures of these capabilities. After examining use of expert judgement on human tests (Chapters 3 and 4), the project is considering the use of more direct tests of AI. In such a model, systems are evaluated on various domains of capabilities and different tasks stemming from these domains.

Alongside benchmarks (Chapter 6), evaluation campaigns for AI and robots are common types of direct tests for AI. Evaluation campaigns refer to a structured and organised effort to assess the performance of AI models or systems using benchmarks or datasets. These campaigns are often organised by research institutions and industry groups as a catalyst for development of these technologies in the last decades. They are central to informing about the maturity of AI and its complementarity to human intelligence.

This chapter provides an overview of the general structure of evaluation campaigns. It lists major campaigns from the National Institute of Standards and Technology (NIST) in the United States and the French Laboratoire National de Métrologie et d'Essais (LNE) in different areas of AI and robotics. It then proposes a method for systemising existing campaigns and identifying tasks unexplored by these evaluations.

The first section explains why a systematic mapping of evaluations of AI and robotics is needed. The chapter then describes a method for mapping evaluation tasks and applies this method to campaigns organised by NIST and LNE. It discusses how to compare evaluated AI capabilities and human skills and presents initiatives evaluating human-AI interaction. Finally, it highlights the limitations of evaluation campaigns and the approach proposed for mapping them.

The need for systematising AI and robotics evaluations

Not all tasks automated by AI and robotics have been evaluated

Many areas for potential AI applications started with challenges too big to be solved and had to be broken into smaller problems. As a result, more evaluation campaigns are concerned with low-level (parsing, recognition, etc.) rather than high-level tasks (automatic speech recognition).

The coverage of high-level tasks is improving with the maturation of intelligent technologies. For example, Deep Fakes Generation rose from scratch and quickly became a subject of societal concern. Deep Fakes can be AI-generated videos staging false events involving real people, such as a speech by Barack Obama that he never gave. As a response to these maturing technologies, prominent AI companies are organising Deep Fake Detection challenges as Kaggle events, an online data science competition where participants use machine learning to solve specific problems.

Beyond individual tasks, these evaluations do not represent all application areas of AI and robotics. This is because large datasets and multiple campaigns are necessary to make a task operational. Moreover, companies usually finance evaluations of tasks only when they have a minimal level of maturity, as well

as commercial potential. Consequently, there might be a gap between AI and robotics capabilities in the academic world and the expectations of industrial actors. This gap often becomes apparent in the choice of evaluation campaigns conducted.

Not all AI tasks are relevant for humans

Evaluation methods often require comparing the output of intelligent systems to reference annotations defined by human experts. However, evaluation tasks are not always relevant for assessing human capabilities. In other words, only some tasks consider human performance as a baseline; having a human-made gold standard does not equate to comparing the performance of AI and humans.

For example, diarisation is considered a building block for more complex speech processing tasks. Automatic speech recognition, for example, transcribes speech or dialogues into text, and includes the identification of each speaker present. Reciprocally, tasks aiming at evaluating (and thus improving) human memory are not really useful for AI.

Not all AI/robotics tasks aim to be entirely independent from humans

Given the many limitations of mechanisms, sensors and algorithms, many tasks automated by robots still require close interaction with human operator(s). Systems that preclude collaboration with humans may be less robust and less effective. For the foreseeable future, designing robotic systems that can team up with humans will leverage the strengths of each: the human's fine dexterity and expert knowledge with the robot's strength and endurance.

Initial efforts are under way to understand how to measure human-robot interaction. As these develop, they can guide the design and implementation of systems. An effective partnership between a human and an AI-based system can also help the AI learn through demonstration and other means.

Framework structure

The framework describing evaluation tasks for AI systems and robots proposed in this study requires identifying key attributes of these tasks.

An evaluation consists in:

1. defining a task to perform
2. presenting a candidate system implementing a function to perform this task with a defined dataset of input
3. measuring the quality of its output (or other characteristics of interest), usually against a dataset of reference.

During these tests, the function of the system itself is considered a black box. Its objective is to transform input data into outputs. A task is independent of the underlying technical components (type of AI algorithm, hardware performing the calculation, etc.). However, it may influence the tests' modalities and environment (e.g. datasets).

Major areas of AI and robotics have been defined from a pairing of these "input data" and "transformation" descriptors. In this paper, they are called "field" and "sub-field". They were included in the mapping to bring out the different classes and families of evaluation tasks.

This document aims at mapping the landscape of tasks that researchers and companies have been trying to automate using AI. To do so, it will consider tasks for which at least one evaluation campaign was devised and resulted in significant progress in the field. This progress could take the form of performance

(i.e. the systems got increasingly closer to solving the task at hand). It could also be measured from a methodological standpoint (i.e. companies or researchers could conclude on how to spur improvement through further campaigns).

Functionality level: High-level vs. low-level tasks

The framework comprises two “functionality levels”: high and low.

High-level describes a task commonly performed by a human that requires some degree of intelligence whose automation can only be brought by an AI system (and not simpler software). An AI system tackling high-level tasks may be used to replace human intervention in a professional setting. It may therefore partially or fully automate certain jobs (a job generally consisting of a set of tasks).

Low-level tasks are intermediate functionalities used to break down a more complex (generally high-level) task into smaller and more manageable problems. The framework in this study specifies the level of each task so it can be more clearly positioned in the perspective of the evolution of work and skill.

Integration level: Pipeline vs. end-to-end systems

High-level tasks are commonly first addressed by *pipeline solutions*. In this case, AI systems consist of a series of sub-systems, each tackling a low-level task to produce the expected high-level output. As the understanding of a problem progresses and AI algorithms evolve, some tasks can be tackled using an *end-to-end* solution. In this case, a single model is learnt to solve the task rather than several specialised modules put together.

Such progress generally comes with substantial performance gains. AI pipelines are hindered by error propagation through the modules and their overall performance cannot exceed that of their weakest component. Therefore, pushing the performance of an AI pipeline forward requires that all components are constantly improved in parallel, which is difficult. It also imposes a more rigid structure. All components play a role partly determined by the roles of the other components. This complicates the emergence of disruptive approaches both at the module-level and the architectural level.

End-to-end solutions alleviate these problems while raising others, such as how the AI system processes the task in an opaque manner. Indeed, it is hard to understand how an end-to-end solution breaks down a task. Assuming such analysis is performed and inefficient steps are identified, it is even harder to make the model more efficient.

On the other hand, end-to-end solutions are typically part of a larger intellectual framework and generally contribute to significant advances in task performance. Machine Translation (MT) is a good example of this evolution. It relied on pipelines for a long time with stagnating performance (or incremental gains), and an increasing complexity of the components, with some labs focusing on a particular component. The introduction of deep neural networks with an end-to-end solution significantly simplified the architecture and improved performance (Diño, 2017^[4]).

Finally, certain low-level tasks are relevant for several high-level tasks, which might not have all achieved transition to an end-to-end solution. Besides, new high-level tasks regularly arise. In these cases, pipeline solutions are generally the more intuitive and successful approaches available. This explains why pushing the performance of low-level tasks may still be justified.

Additionally, moving from pipeline to end-to-end does not mean the task is solved or becomes easier. However, achieving an end-to-end architecture is a significant milestone in the improvement of AI performance on a task. This is why the framework specifies the task integration level.

Comparison with human capabilities

Typically, evaluation campaigns use human performance as a reference. However, human capabilities themselves are not necessarily easy to quantify or generalise. For example, researchers have argued that the Machine Learning (ML) community “has lacked a standardized, consensus framework for performing the evaluations of human performance necessary for comparison” (Cowley et al., 2022^[5]).

Comparative results with human participants “should be approached with caution: when human factors, psychology, or cognitive science research experts, and experts in other fields that study human behaviour scrutinize the methods used to evaluate and compare human and algorithm performance, claims that the algorithm outperforms human performance may not be as strong as they originally appeared” (Strickland, 2019^[6]).

Even with the noted deficiencies, benchmarking based on human-curated datasets is the foundation upon which the stunning progress in AI has been built. Some benchmarks have been “saturating”, with ML algorithms achieving parity/near-parity with human performance at increasing speed (Thrush et al., 2022^[7]). This creates a necessity for finding efficient means of updating, extending and diversifying benchmark data. Efforts to address this need are emerging, such as Dynatask (Thrush et al., 2022^[7]).

To establish a bridge between AI and human capabilities, this study considers a “human similarity level” of performance for AI tasks. Such a reference creates a direct and clear link between AI and human abilities for a strictly defined context (i.e. the AI task), making for a straightforward comparison tool. For high-level tasks, comparing human and AI performance helps understand how an AI solution can substitute for human labour in the given task. For low-level tasks, it highlights the bottlenecks of pipeline solution and may give insight into which human abilities and skills are hard to automate.

As an added advantage, comparison based on human performance provides insights on the intrinsic difficulty of the task. Some tasks tackled by AI research are difficult even for humans, an important factor when considering the performance of an AI solution. In addition, human-level performance remains in many cases the highest reachable standard (although in some tasks, AI does outperform humans).

Arguably, many tasks that embodied artificially intelligent systems, such as robots, may try to perform are easy for humans. A classic example is a chess-playing robot. AI has produced systems that perform better than even most grandmasters. However, it is still challenging for a robotic system to pick up and move chess pieces reliably under uncontrolled ambient lighting and other real-world conditions. Yet even a young child can pick up a piece they’ve never seen before from the middle of a board and place it in a new square without colliding with other pieces or dropping it.

The method in this document to estimate the human performance level varies with the tasks, the learning paradigm and the evaluation settings. Many evaluation campaigns provide gold standard annotations for supervised learning. In this case, the human performance level is by definition 100%. This is because humans usually make the gold standard and all corner cases have been removed. If a human is not sure, the example cannot reasonably be used to train an AI system.

Evaluation campaigns of AI capabilities

Evaluation campaigns at LNE, NIST and other institutions

NIST and LNE have supported the advancement and implementation of emerging technologies such as AI and robotics through the development of measurement science. Measurement science encompasses the identification of performance requirements for a given task or domain, definition of metrics for the performance requirements and development of evaluation infrastructure. The evaluation infrastructure may include test methods, test artefacts, datasets, testbeds and other tools.

Box 7.1. Facet characteristics of the LNE and NIST evaluations vs. those of benchmark tests

To illustrate how the characteristics of AI evaluation campaigns compare to the characteristics of AI benchmark tests, we use the facet indicators from Chapter 6 to describe eight of the NIST and LNE evaluation campaigns. Neither the benchmark tests discussed in Chapter 6, nor the evaluation campaigns discussed in this chapter, were selected according to the facet values.

Figures 7.1 to 7.3 show the frequencies of the different values from the 18 facets, in a similar manner to Figures 6.2, 6.3 and 6.4 in Chapter 6. Labels appearing in green bold represent the desirable values, referring “to the preferred or most challenging case”. For the complete evaluations and attribution of the facets to the different campaigns, see Annex 7.B.

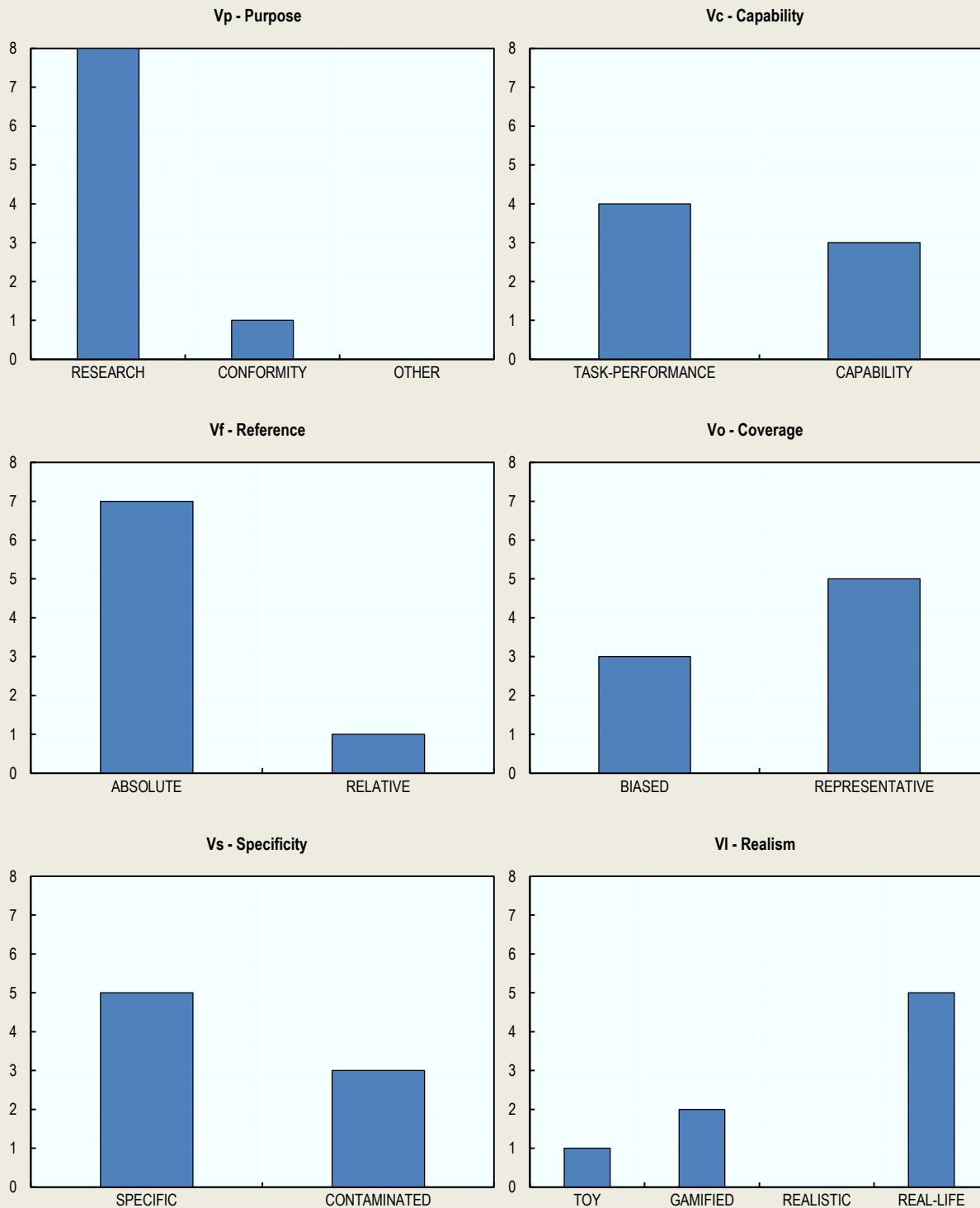
With regards to the **Validity** group (“Does it measure what we want to measure?”) and illustrated by the first six graphs, five of six facets (i.e. Purpose, Capability, Reference, Coverage and Specificity) have frequencies similar to the 36 benchmarks from the previous chapter (see Figure 7.1). The major difference arises from the Realism facet, for which the eight campaigns are evaluated as being more real-life than realistic instruments, as is the case for the 36 benchmarks. Both groups of instruments display the desirable values for Reference, Coverage and Specificity facets, whereas the Capability facet is still underrepresented.

Concerning the **Consistency** group (“Does it measure it effectively and verifiably?”) and illustrated by the next six graphs, there are more differences across facets between the two evaluation exercises (see Figure 7.2). Among the 36 benchmarks, more are found to have *automated* Judgeability, *exact* Reproducibility and to be *Reliable* compared to the eight evaluation campaigns. These three values represent the preferred values of the facet. This suggests that the 36 benchmarks evaluated in the previous chapter are overall more consistent than the eight evaluation campaigns by NIST and LNE. This is a plausible result stemming from the contrast between one-time evaluations adapted to the application needs of specific sponsors – which can rely on evaluations specialised to their applications – and the more general focus of most benchmarks.

Finally, regarding the **Fairness** group (“Does it treat all test takers equally?”) and illustrated by the last six graphs, no clear difference in the frequencies is found between the two groups and across the facets (see Figure 7.3). Both groups display the preferred values from the Ambition, Objectivity and Autonomy facets, suggesting an appropriate fairness for these instruments.

Overall, there is a similar pattern of attributed facet values for the evaluation campaigns and the benchmarks.

Figure 7.1. Rater values selection on validity facets for eight evaluation campaigns by NIST and LNE




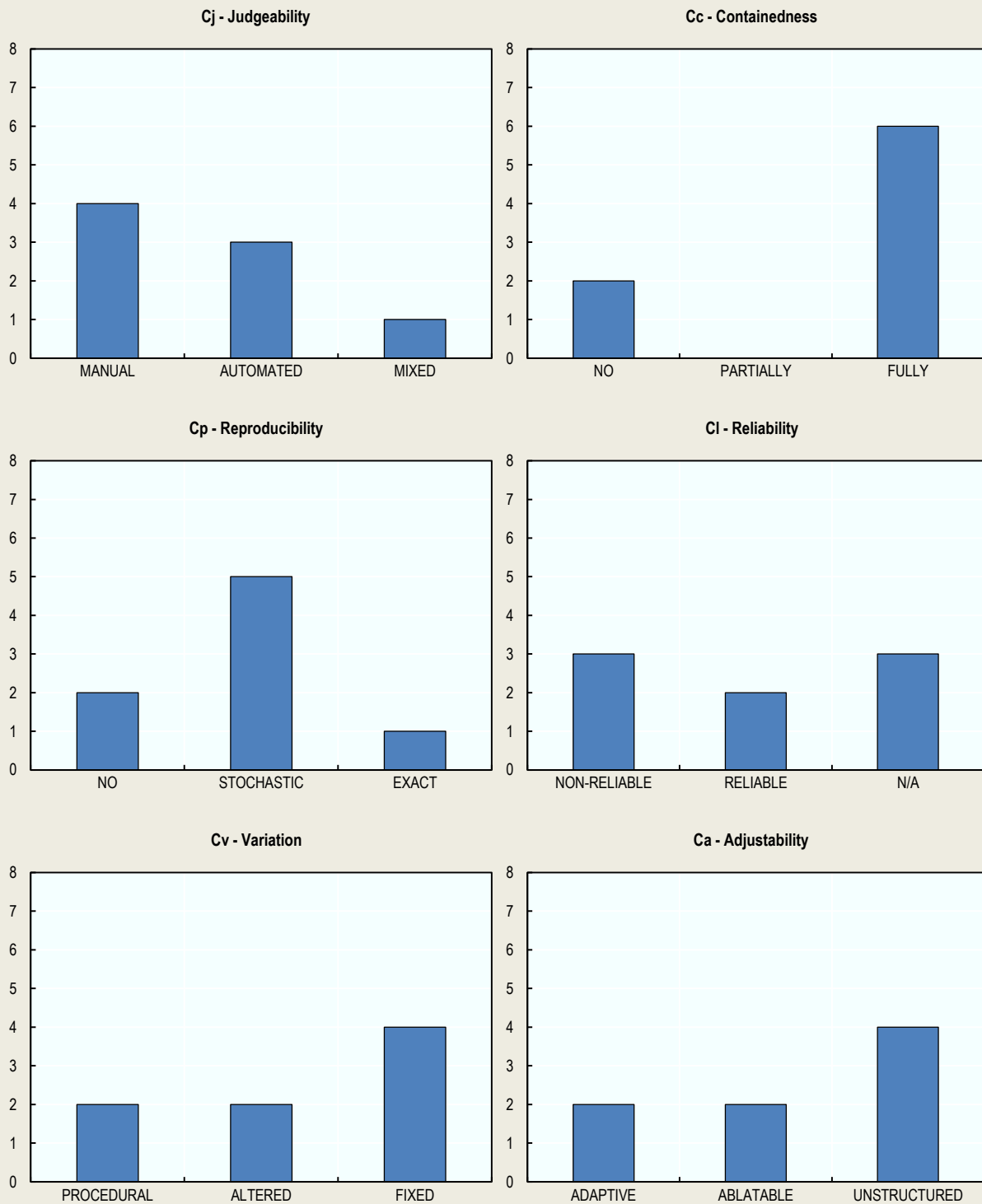
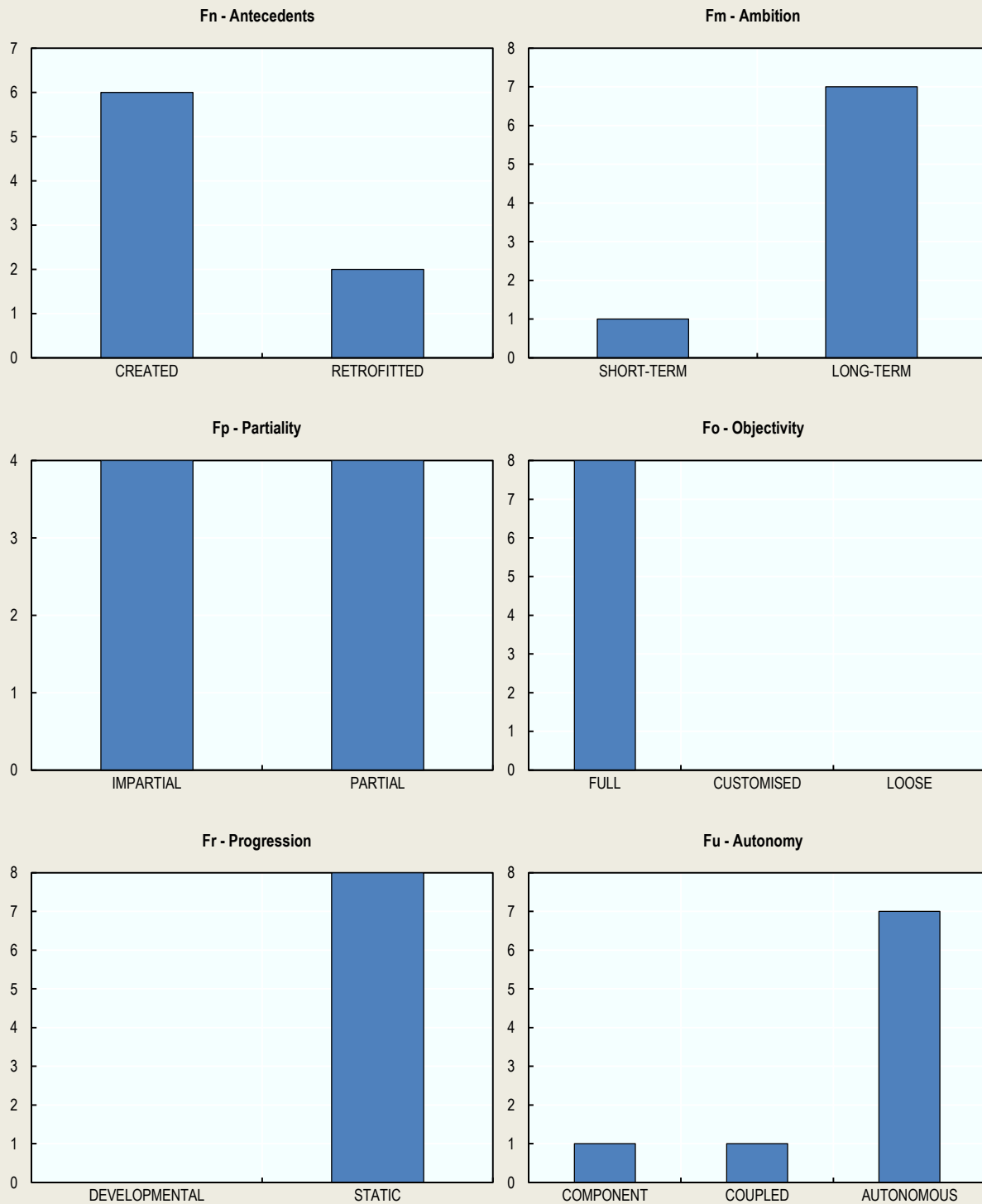
StatLink  <https://stat.link/snb480>

Figure 7.2. Rater values selection on consistency facets for eight evaluation campaigns by NIST and LNE



StatLink  <https://stat.link/xi51km>

Figure 7.3. Rater values selection on fairness facets for eight evaluation campaigns by NIST and LNE



StatLink  <https://stat.link/619s5r>

The approach for developing the evaluation infrastructure depends on the community's needs, the maturation trajectory of the technologies to be evaluated and other considerations. Therefore, adaptability is always required. For instance, some projects address evaluation needs through recurring competitions that increase the complexity and difficulty in each iteration. Others take place within standards development organisations. New domains and technologies are tackled based on the needs and priorities of industry and academia and proceed in collaboration with stakeholder communities. Over the years, NIST and LNE each carried out hundreds of system evaluations and evaluation campaigns.

The rest of this section presents part of this work and breaks down AI into three major fields – natural language processing (NLP), computer vision and robotics – sub-fields and tasks, each exemplified by one or more evaluation campaigns, including Evaluation en Traitement Automatique de la Parole (ETAPE) (Galibert et al., 2014^[8]), REPERE (Kahn et al., 2012^[9]), Moyens AUTomatisés de Reconnaissance de Documents ecRits (MAURDOR) (Brunessaux et al., 2014^[10]) and FABIOLE (Ajili et al., 2016^[11]).

As mentioned in the previous section, high-level tasks offer a closer comparison to human capabilities necessary for work. This is in line with the AIFS project's desire to compare AI capabilities to those of humans. As a result, this section only discusses high-level tasks; lower-level tasks are illustrated in the Annex 7.A. Finally, it discusses adjustments to evaluation protocols needed by NIST and LNE to allow the evaluation of multiple AI and robotics solutions, as well as some of their comparison to human performance.¹

Natural language processing

NLP is the field of AI that enables computers to process and produce human language. Language can be conveyed via several media, the most common being text and speech. This is reflected in NLP, where text and speech processing are two separate sub-fields. As language is a major medium for communication, NLP intertwines with many other AI fields.

Text processing and text comprehension

Text processing and text comprehension are the sub-fields of AI focusing on enabling computers to interact with humans using text. It is a fundamental stake of AI to communicate through language, especially text, as it is the most natural way that most humans communicate. However, language is fundamentally fuzzy, making it particularly challenging. Table 7.1 presents a number of high-level tasks from these sub-fields and examples of related evaluation campaigns. Lower-level tasks, such as named entity recognition, story segmentation or extraction of relations between textual phrases, are found in Annex 7.A.

NLP tasks are heavily influenced by their textual context, defined by the domain (field of knowledge) and genre (type of text, such as tweets or articles). Genres are not hierarchical, which means that proficiency in one genre does not guarantee efficiency in others. This context specificity suggests that a universal NLP approach is elusive. Evaluating AI system performance in NLP thus requires clear concept definitions, crucial for creating annotation schemas and evaluation protocols.

The case of the topic detection and tracking (TDT) task – which suffered from vaguely defined central concepts and was interrupted after only three iterations – exemplifies this need for clear definitions. TDT's rapid cessation also stemmed from its ambitious goals that overlooked the existing state of the art. The domain's significance is also illustrated in the development of conversational agents and question-answering systems. For instance, these systems rely heavily on domain-specific knowledge and the kind of response expected, be it closed, factual, list-based or open answers.

Table 7.1. Text processing and comprehension high-level task examples and associated evaluation campaigns

Task	Nature of the input data (application area)	Evaluation campaign name (organisers)	Timespan	Integration level	Difficulty
<i>Conversational agent (chatbot)</i>	Culinary recipes	LIHLITH (LNE)	2020-2022	Pipeline	~60% success rate for task-oriented systems, it drops <30% for open dialogue.
<i>Topic detection</i>	Newswires	Topic Detection and Tracking (TDT) (NIST)	1997-2004	End-to-end	70% < success rate < 95% depending on the type of data.
<i>Topic tracking</i>	Newswires	Topic Detection and Tracking (TDT) (NIST)	1997-2004	End-to-end or pipeline, depending on the use case	~60% success rate.
<i>Question-answering</i>	Web content	QUAERO (LNE)	2008-2014	End-to-end	Success rate ~60%, with variability due to application domains and metrics used.
	QA by smartphone personal assistant	INC (LNE)	2019		
<i>Machine translation</i>	Newspaper articles and broadcast news transcriptions from various radio and television programmes, blog articles, useNet pages, mails	QUAERO (LNE), TRAD (LNE)	2009-2014 2012-2014	End-to-end	Success rate is ~35%, but the metric is extremely punishing since only one correct target translation is considered. Human evaluation is more forgiving and displays performance level > 70%.
	Newspaper	MT (NIST)	2001-2015		

As mentioned previously, evaluation campaigns also face the challenge of AI and human comparison. To address the drawbacks of current methods used to evaluate MT technology, NIST initiated a meta-campaign, Metrics for Machine Translation Evaluation (MetricsMaTr)². It noted the following drawbacks:

- Automatic metrics have not yet been proven able to predict the usefulness and reliability of MT technologies with respect to real applications with confidence.
- Automatic metrics have not demonstrated they are meaningful in target languages other than English.
- Human assessments are expensive, slow, subjective and difficult to standardise.

The MetricsMaTr evaluation tests automatic metric scores for correlation with human assessments of MT quality for a variety of languages, data genres and human assessments.

Speech processing

Speech processing focuses on all tasks allowing a computer to understand and produce speech (Table 7.2). Lower-level tasks, such as diarisation, language identification or story segmentation are found in Annex 7.A.

Similar to text processing, speech processing is a fundamental field for man-machine interaction, and thus at the crossroads of many scientific domains. For instance, speaker verification systems are designed to determine whether a specific audio segment was spoken by a particular individual. These systems are especially useful in forensic applications. For example, the voice of a suspect might need to be identified despite noise or signal distortions.

Whereas speaker verification focuses on confirming if a specific individual voiced a given segment, speaker recognition pinpoints all instances a particular person speaks across various audio clips. To

assess its effectiveness, the system is presented with texts or audio, and its outputs are measured using predetermined metrics. Accuracy can be influenced by the availability and quality of audio samples and any potential noise or interferences present in them.

Table 7.2. Speech processing high-level task examples and associated evaluation campaigns

Task	Nature of the input data (application area)	Evaluation campaign name (organisers)	Timespan	Integration level	Difficulty
<i>Speaker recognition</i>	Audio debate	QUAERO ¹ (LNE), REPERE (LNE)	2009-2014 2010-2014	Pipeline	~97% success rate. Noisy input may significantly affect performance.
	Forensics, conversational telephone speech	Speaker Recognition (NIST)	1996-2021		
<i>Speaker verification</i>	Audio debates (criminalistics), police student interviews	VOXCRIM (LNE)	2017-2022	Pipeline	90% < success rate < 97% depending on the type of input.
	Forensics, conversational telephone speech	Speaker Recognition (NIST)	1996-2021		
<i>Automatic speech recognition</i>	Smartphone and pad personal assistant	INC (LNE)	2019	End-to-end	75% < success rate < 97% depending on the type of speech and the noise level.
	Audio broadcast news, conversational telephone speech, meeting room speech	Rich Transcription (NIST)	2003-2009		
	Conversational telephone speech	Conversational telephone recognition (NIST)	2019-2021		
	Audio broadcast news	Broadcast news recognition (NIST)	1996-1999		
<i>Information retrieval</i>	Audio broadcast news	Spoken document retrieval (SDR) (NIST)	1997-2000	Pipeline (ASR + text IR)	~65% success rate on English resources. Other languages may exhibit more variability.
<i>Topic detection</i>	Audio broadcast news	Topic Detection and Tracking (NIST)	1998-2004	Pipeline	30% < success rate < 70%.
<i>Topic tracking</i>	Audio broadcast news	Topic Detection and Tracking (NIST)	1998-2004	End-to-end or pipeline, depending on the use case	~70% success rate.
<i>Question-answering</i>	Robot facing a human (assistive robotics)	HEART-MET (LNE)	2020-2023	Pipeline (ASR + Text QA)	50% < success rate < 80%.

Note: HEART-MET, RAMI, ACRE and ADAPT are AI and robotics competition associated with the METRICS project (Avrin et al., 2020_[12]), co-ordinated by LNE (<https://metricsproject.eu/>). ¹ <http://www.quaero.org/>, see also (Ben Jannet et al., 2014_[13]; Bernard et al., 2010_[14])

Speech processing faces challenges similar to text processing, influenced by conversation specificity and discourse construction. Sociolinguistic factors, such as education level, politeness, accent and prosodic markers, necessitate specialised systems. Issues like code-switching, sociolects and varying noise levels also complicate evaluation.

NIST's recent OpenASR21 Challenge tasks or Speaker Recognition Evaluation in 2021 attempt to address these challenges. For instance, OpenASR21 Challenge tasks have more case-sensitive evaluations, and the 2021 Speaker Recognition Evaluation use publicly available corpuses and non-speech audio and data (e.g. noise samples, room impulse responses and filters).

Computer vision

Computer vision (CV) is an AI field focused on enabling a computer to extract information from images and videos, which are another major media for human communication. CV applications are therefore multiple – from automatically processing bank cheques to indexing vast amounts of visual data.

Recognition

Recognition is the sub-field of CV specialising in images (i.e extracting information from a single, fixed image). Table 7.3 presents the high-level image segmentation task from this sub-field and examples of related evaluation campaigns. Lower-level tasks, such as image classification, shape recognition or pose estimation, are found in Annex 7.A.

Table 7.3. Recognition high-level task example and associated evaluation campaigns

Task	Nature of the input data (application area)	Evaluation campaign name (organisers)	Timespan	Integration level	Difficulty
Image segmentation	Administrative documents	MAURDOR (LNE)	2011-2014	End-to-end	60% < success rate < 90% depending on the type of objects, for semantic segmentation. ~40% success rate for instance segmentation.
	Aerial images	MMT (LNE)	2020	End-to-end	

Image segmentation is an example of an advanced procedure within the realm of object detection. Rather than creating a general bounding box around an identified object, this method precisely traces the object's contour. Image segmentation bifurcates into two primary categories: semantic segmentation (where objects of identical classes are uniformly categorised) and instance segmentation (which provides distinct identification for each object within a class).

Such meticulous identification is paramount in contexts that demand precision beyond the capabilities of bounding boxes. These include for the accurate location of specific items or the detailed analysis of medical imagery. The effectiveness of image segmentation techniques is often measured using the Jaccard index, assessing the congruence between predicted and observed segments (Costa, 2021^[15]). Key determinants influencing this procedure include the nature of the objects, their positioning and ambient environmental conditions, such as illumination.

Motion analysis

Motion analysis is the sub-field of CV specialising in the analysis of video feeds. Video feeds propose specific challenges and thus specific tasks. However, these can also be considered a special case of application for recognition tasks (with the temporal component implying a continuity constraint). Therefore, many recognition tasks are also explored in a video setting. For clarity and concision, the recognition tasks carried in a video setting are not re-introduced. Table 7.4 presents a number of high-level tasks from this sub-field and examples of related evaluation campaigns. Another aspect of the lower-level shape recognition task is presented in Annex 7.A.

Face recognition, an example of this sub-field, involves systems using biometrics to analyse facial features and associate them with specific identities, like first and last names. This technology is instrumental in security applications, such as access control. It is also employed for automatic video indexing by identifying celebrities and TV hosts. Beyond this, it is leveraged to enhance tasks like speaker recognition, as demonstrated in the REPERE campaign.

Table 7.4. Motion analysis high-level task examples and associated evaluation campaigns

Task	Nature of the input data (application area)	Evaluation campaign name (organisers)	Timespan	Integration level	Difficulty
<i>Optical Character Recognition (OCR)</i> <i>Face recognition</i>	Multimodal television streams	REPERE (LNE)	2010-2014	End-to-end	~85% success rate.
				End-to-end	> 99% success rate. In some conditions, algorithms have performed better than humans.
<i>Object detection</i>	RGB camera feed from a fixed angle (logistics robotics)	BLAXTAIRSAFE (LNE)	2019	End-to-end	30% < success rate < 90% depending on the type of data (environmental conditions, types of object, etc.).
	RGB camera feed from a robot (assistive robotics)	HEART-MET (LNE)	2020-2023		
	Underwater and aerial RGB camera feeds from robots (inspection & maintenance robotics)	RAMI (LNE),	2020-2023		
	RGB camera feed from a fixed angle on a workbench (agile production robotics)	ADAPT (LNE)	2020-2023		
<i>Tracking</i>	Industrial parts	E3064-16 Standard Test Method for Evaluating the Performance of Optical Tracking Systems that Measure Six Degrees of Freedom (6DOF) Pose (NIST), E3064-16 Standard Test Method for Evaluating the Performance of Optical Tracking Systems that Measure Six Degrees of Freedom (6DOF) Pose (NIST), E3124-17 Standard Test Method for Measuring System Latency Performance of Optical Tracking Systems that Measure Six Degrees of Freedom (6DOF) Pose (NIST)	2016-present	Pipeline	60% < success rate < 80%. A difficulty of tracking is to continuously assign a bounding box to the tracked object (or to continuously predict its correct contour). In this task, accuracy of identification is usually good, but the overlap over time between the system's bounding box and the reference is poor.
<i>Image segmentation</i>	Robot camera streams (agricultural robotics)	ROSE ¹ (LNE)	2018-2022	Pipeline	60% < success rate < 90% depending on the type of data.
		ACRE (LNE)	2020-2023		
	Camera feed from a fixed angle (agile production robotics)	ADAPT (LNE)	2020-2023		
	Multimodal television streams	REPERE (LNE)	2010-2014		
<i>Shape recognition</i>	Underwater robot camera feed (underwater inspection and maintenance robotics)	RAMI (LNE)	2020-2023	End-to-end	~85% success rate.
<i>Information retrieval</i>	General domain videos (IACC.3), Vimeo clips (V3C1 dataset)	TRECVID (NIST) - Ad hoc Video Search (AVS)	2001-present	Pipeline (combines several CV)	Success rate <60% with strong variation between systems.

Task	Nature of the input data (application area)	Evaluation campaign name (organisers)	Timespan	Integration level	Difficulty
	BBC rushes, BBC Eastenders, Flickr videos	TRECVID (NIST) - Instance Search (INS)		modules such as OCR and face recognition + multimodal systems such as speech processing and/or NLP)	Success rate <15%.
	Aerial images (LADI dataset + NIST dataset)	TRECVID (NIST) - Disaster Scene Description and Indexing (DSDI)			Success rate <40%.
	Vines videos, Vimeo clips (V3C2)	TRECVID (NIST) - Video to Text Description (VTT)			Success rate <60% with strong variation between systems.
	Outdoor surveillance footage	TRECVID (NIST) - Activities in Extended Video (ActEV)			Success rate <60% with strong variation between systems.
Video summarisation	BBC Eastenders	TRECVID (NIST) - Video Summarisation (VSUM)		End-to-end	50% < success rate < 60%.

Note: ¹Challenge ROSE (RObotique et capteurs au Service d'Ecophyto): <http://challenge-rose.fr/>

Typically, the evaluation of face recognition is approached as a binary classification task. The system's accuracy can be affected by factors like facial orientation and lighting conditions. For instance, in the REPERE campaign, the evaluations used high-quality video segments from TV shows that featured optimal lighting and well-composed shots of individuals.

Regarding challenges in CV, the data selected profoundly determine the system's capacity and reach. Environmental factors, such as lighting, distance and backdrop, play crucial roles, especially in tasks like action recognition. Enhancing system robustness often mandates a new dataset, but optimised performance isn't always retained across datasets. Some tasks, like vision for autonomous driving, confront inherent noise. Here, traditional RGB cameras may be supplemented with infrared cameras or Light Detection And Ranging (LiDAR) to discern depth and navigate challenging conditions. Evaluations typically compare like-to-like image categories, and outcomes might not reflect the system's efficacy with poor-quality images. The E1919-14 standard gauges a static optical system's performance under strictly controlled conditions, which may not represent real-world application performance.

Robotics

NIST and LNE develop measurement infrastructure for evaluating robots used in emergency response and industrial/manufacturing applications. These robotic systems can be considered examples of embodied AI. Robot evaluations cover the range of functionality levels – from basic “competences” such as vision or image processing through high-level whole system task performance. For instance, industrial robot benchmarking (Norton, Messina and Yanco, 2021_[16]) categorises evaluations as mobility, manipulation, sensing or interaction. Table 7.5 and Table 7.6 focus on locomotion (mobility) and manipulation, building upon the sensing discussed above related to CV and some of the interaction algorithms, such as for chatbots.

Efforts are emerging on evaluation of human-robot interaction but are not mature. The Institute of Electrical and Electronic Engineers (IEEE) has launched a study group on human-robot interaction metrics. It has begun developing foundations for standards, such as recommended practices for human-robot interaction design of human subject studies. An overview of potential approaches for evaluation of human-robot interaction (HRI) can be found in Marvel et al. (2020_[17]).

Locomotion

Locomotion, a sub-field of robotics, allows a robot to move in its environment. This is a key skill for autonomous robots. Table 7.5 presents a number of high-level tasks from this sub-field and examples of

related evaluation campaigns. Lower-level tasks, such as balancing, swimming or arial navigation, are found in Annex 7.A.

Evaluation campaigns for robotics span a broad spectrum of scenarios and standards, each designed to assess specific capabilities while accounting for the complexities of real-world interactions. Some standards, like the ASTM E2826/E2826M-20³, focus on how robots move across specific terrains. Others like ASTM F3244-21 evaluate navigation capabilities but only within areas with static obstacles. These standards, even while being comprehensive, are sometimes restricted in their scope. For example, the ASTM F3499-21 mainly assesses a vehicle's precision in aligning with a docking location. It does not delve deeply into other aspects of the docking process.

Table 7.5. Locomotion high-level task examples and associated evaluation campaigns

Task	Nature of the input data (application area)	Evaluation campaign name (organisers)	Timespan	Integration level	Difficulty
<i>Walking</i>	Stepping over stones, on a beam, on flat ground, on a slope, over obstacles.	ROBOCOM++ (LNE)	2017-2021	Pipeline	80-100% depending on the task and temperature.
	Tests were performed in a climatic chamber through a range of temperatures.				
<i>Stairs</i>	Climbing stairs (10 cm high stairs without handrail, 15 cm high stairs with handrail). Tests were performed in a climatic chamber through a range of temperatures.	ROBOCOM++ (LNE)	2017-2021	Pipeline	40-100% depending on temperature.
<i>Crossing harsh terrains</i>	Traversing sandy, rocky terrains, moving through indoors structure with debris, crossing gaps, hurdles, traversing at sustained speed.	ASTM E54.09 Standard Test Method Suite for Evaluating Robot Mobility (NIST). Individual test methods for:	current	Pipeline	Varies by type of terrain. Most implementations include humans-in-the-loop. Estimate that autonomous implementations average 50% success at most across all types.
		- terrain types: flat wood, sand, gravel; crossing pitch/roll, continuous pitch/roll, step fields			
		- obstacle types: variable hurdles, variable gaps			
		- various stair types			
<i>Rolling</i>	Over roads, agricultural plots or indoor environments. Location precision and speed are measured.	3SA (LNE)	2020-2023	Pipeline	80-100% performance rate.
		ROSE (LNE)	2018-2022		
		ACRE (LNE)	2020-2023		
		HEART-MET (LNE)	2020-2023		
<i>Flying</i>	Flying in a known environment (industrial site imitation).	RAMI (LNE)	2020-2023	Pipeline	Good performance but susceptibility to wind.
<i>Navigation</i>	Underwater navigation and mapping without Global Navigation Satellite System (GNSS), with added passive beacons.	RAMI (LNE)	2020-2023	Pipeline	Poor performance, slow.
	Inside a warehouse with defined and undefined structured and unstructured areas.	F3244-17 Standard Test Method for Navigation: Defined Area (NIST)	current		Many systems can succeed but it is configuration-dependent (both of the test course and the robot).

Task	Nature of the input data (application area)	Evaluation campaign name (organisers)	Timespan	Integration level	Difficulty
<i>Area covering</i>	Area disinfection with UV lamp (assistive robotics).	HEART-MET (LNE)	2020-2023	Pipeline	80-100% success rate but poor performance (slow).
<i>Docking</i>	Navigating inside a warehouse.	F3499-21 Standard Test Method for Confirming the Docking Performance of A-UGVs (NIST)	current		Data on success rates not yet available.
<i>Avoiding unexpected obstacle on course</i>	Navigating inside a warehouse.	F3265-17 Standard Test Method for Grid-Video Obstacle Measurement (NIST)	current		Some systems detect and react quickly enough, but not all.

Evaluations of autonomous cars offer a good comparison between AI and human performance. For instance, autonomous cars are deployed in certain states and cities to collect data and improve their performance and safety. In the United States, the National Highway Traffic Safety Administration (NHTSA) requires manufacturers and operators to report crashes involving vehicles equipped with Automated Driving Systems Society of Automotive Engineers (SAE) levels 3 through 5. There are not much data yet, but 130 crashes were reported between July 2021 and 15 May 2022. Data on the number of vehicles or number of miles driven are not required. Therefore, it is hard to compare self-driving vehicle safety performance to human-driven cars. NHTSA reports 6 102 936 crashes in 2021, with a projected rate of 1.37 fatalities per 100 million vehicle miles.

A study by Virginia Tech's Transportation Institute (Blanco et al., 2016^[18]) from 2016 compared estimated crashes for human-driven versus autonomous vehicles and found that autonomous vehicles have a lower crash rate, especially when it comes to severe crashes. Moreover, crash rates in the Second Strategic Highway Research Program (SHRP 2) National Driving Study (NDS) dataset surpassed those of autonomous vehicles across all severity levels (see Figure 1 in Blanco et al. (2016^[18])). There has apparently been no follow-up work to update these estimates from 2016.

Manipulation

Manipulation refers to a robot's ability to interact with its environment using effectors, typically robotic arms and hands (or grippers). Robotic manipulation is crucial in various industries. In manufacturing, robots perform repetitive tasks, while in health care they might assist in surgeries. The challenges in this domain often revolve around dexterity, adaptability to different objects and environments, and the integration of sensory feedback for more nuanced and delicate operations. Table 7.6 presents a number of high-level tasks from this sub-field and examples of related evaluation campaigns. Lower-level tasks, such as picking-and-placing, handing an object over and pouring are found in Annex 7.A.

As mentioned, these evaluations offer pivotal insights into the various capabilities of robots. However, they also come with certain limitations. For instance, the ROSE Challenge centres on the weeding of particular plants. While it offers a controlled environment by manually sowing weeds, the task becomes intricate due to unpredictable furrows and environmental conditions. This makes it challenging to manoeuvre robots and identify weeds. The controlled nature of evaluations must be offset against the unpredictable variables of real-world applications.

As mentioned, these evaluations offer pivotal insights into the various capabilities of robots. However, they also come with certain limitations. For instance, the ROSE Challenge centres on the weeding of particular plants. While it offers a controlled environment by manually sowing weeds, the task becomes intricate due to unpredictable furrows and environmental conditions. This makes it challenging to manoeuvre robots and identify weeds. The controlled nature of evaluations must be offset against the unpredictable variables of real-world applications.

Table 7.6. Manipulation task examples and associated evaluation campaigns

Task	Nature of the input data (application area)	Evaluation campaign name (organisers)	Timespan	Integration level	Difficulty
<i>Weeding</i>	In tests performed with maize and bean and several types of weeds, performance is measured by amount of remaining weeds and damaged crops.	ROSE (LNE)	2018-2022	Pipeline	20-80% success rate but slow and dependent on crop growth and weather conditions.
<i>Task-oriented grasping</i>	Grasping sleds and crossing terrain.	E2830-11(2020) Standard Test Method for Evaluating the Mobility Capabilities of Emergency Response Robots Using Towing Tasks: Grasped Sleds (NIST)	current		Low success when run fully autonomously.
<i>Assembly</i>	Peg insertions, gear meshing, electrical connector insertions, nut threading.	Assembly task board 1 (NIST)	current		Dependent on set-up. Many seem to pre-programme carefully, in which case success is higher. Full autonomous success is estimated to be <20%.
	Alignment and insertion of collars and pulleys, handling flexible parts, meshing/threading belts, actuating tensioners and threading bolts.	Assembly task board 2 (NIST)	current		Dependent on set-up. Many seem to pre-programme carefully, in which case success is higher. Full autonomous success is estimated to be <10%.
	Tracking, placement, weaving and manipulation of loose cables, handling flexible parts and inserting ends into various connectors.	Assembly task board 3 (NIST)	current		Dependent on set-up. Many seem to pre-programme carefully, in which case success is higher. Full autonomous success is estimated to be <10%.
<i>Manipulating object mobile parts</i>	Opening cupboards and drawers (assistive robotics).	HEART-MET (LNE)	2020-2023	Pipeline	Poor success rate and performance.
	Opening valves (underwater inspection and maintenance robotics).	RAMI (LNE)	2020-2023		

In terms of manufacturing robot performance versus humans, the designs of the NIST task boards for benchmarking small parts assembly are inspired by the classification tables in the Boothroyd-Dewhurst design-for-assembly method (Boothroyd, Dewhurst and Knight, 2010_[19]), which can be used to estimate human performance. For instance, for an early variant with simple peg-in-hole insertion tasks, the classification tables yield an estimated completion time of 2.5 seconds.

Several factors complicate direct comparison with human performance in industrial settings. These include robot programming/teach time, and trade-offs regarding how unsafe or dull tasks may be for humans. However, NIST assembly task boards present challenges in specific tasks like peg insertions, gear alignments and handling of flexible components. Some teams rely on traditional methods, such as lead-through programming, and most tasks are done in simpler horizontal configurations.

Another comparison is with robot versus human performance in assembly-related operations (using elements from the NIST assembly task boards) and based on deep reinforcement learning. Luo et al., (2021_[20]) evaluated the hand-eye co-ordination of a robot trained for 12 hours to insert an HDMI plug into a moving receptacle. The robot's performance was comparable with that of humans (see Fig. 8 in Luo et al., (2021_[20])).

Going beyond “basic” interaction

This section discusses evaluation initiatives of tasks closer to human capabilities (such as reasoning, emotion perception or human interaction). This contrasts with more “basic” tasks of perception or interaction, e.g. speech recognition, text understanding, object recognition or object grasping. The term “initiatives” here describes all manners of evaluation on a scale larger than a few systems benchmarked in a single research paper. These somewhat high-level tasks rest on the more “basic” tasks that form the bulk of AI research; it remains challenging to obtain high performance with proper robustness.

Efforts are emerging to foster innovations in metrology for effective, real-world HRI. HRI is a vast and interdisciplinary area of study that has lacked cohesion and even a common vocabulary. NIST has been collaborating with several international researchers to begin developing consensus on metrics along with repeatable and reproducible HRI research. Bagchi et al. (2022_[21]) identified areas being pursued following workshops with stakeholders:

- guidelines for reproducible and repeatable studies with quantifiable test methods and metrics
- human dataset creation and transferability of such content
- a central repository for hosting such datasets, as well as software tools for HRI
- standards of practice for HRI, particularly for human studies.

Marvel et al. (2020_[17]) define a comprehensive framework and test methodology for the evaluation of human-machine interfaces and HRI, with a focus on collaborative manufacturing applications. Their framework encompasses four levels of human-robot collaboration to be examined – from total separation to supportive and simultaneous work on a same workpiece to complete a common task.

A comprehensive framework must include verbal, non-verbal and other cues, as well as measures of a human-robot team’s effectiveness. While studies have measured effectiveness, user experience and other factors, there are no benchmarks for this domain. The IEEE initiated a new standards study group on metrology for HRI in 2021⁴.

Rapidly developing areas of concern and study related to AI involve risk, bias, trustworthiness and explainability. NIST has begun laying the groundwork to develop work in these and related areas. Metrics and evaluation methods are anticipated.

Overall, more complex tasks form niche communities with slow development, which in turn produces low need for a strong evaluation framework. An in-between solution is called shared task. This is a regular gathering of the community (usually at a major annual conference) around a common task and a common dataset. One NLP shared task – FinCausal – looks at causal inference and detection in financial texts through two tasks: binary classification and relation extraction (Mariko et al., 2020_[22]).

Shared tasks help structure a community with common evaluation protocols (i.e. tasks, datasets and metrics). However, they tend to have a narrow scope due to organisational limitations. This motivates the multiplication of parallel propositions, all bringing diversity but remaining limited in their scope.

This limited scope is visible in Multimodal Emotion Recognition, with the Audio-Visual + Emotion Recognition (AVEC) challenge (Povolný et al., 2016_[23]). This was the precursor that spawned the Multimodal Emotion Recognition (MEC) Challenge for Chinese language (Li et al., 2016_[24]), among other similarly specialised settings. The development of systems working across tasks is left to the candidates’ initiative, which can slow down integration and the overall maturation of the field. Other tasks were investigated, such as automated reasoning, but were in such early stages of development that shared tasks could not be found.

On a different note, BIG Bench (Srivastava et al., 2022_[25]) is a large NLP benchmark with 204 tasks. BIG Bench evaluates the large language models that form the backbone of most state-of-the-art NLP approaches, such as GPTx. Thus, the datasets associated with each task are small (i.e not enough to train

a large model from scratch), but sufficient to fine-tune these models. Tasks range from solving mathematical problems to answering college-level geography tests. Moreover, a strong human baseline has been established for reference against the models. It is shown that all models perform similarly, with performance improving linearly with the number of parameters of the models. Evaluation also shows that all models perform poorly compared to human performance, with humans averaging around an 80% success rate and models not reaching 20%.

The benchmark has been designed to be hard, thus having a large progression margin. It involved 444 authors from more than 100 institutions, highlighting the potential community that could be structured around these initiatives. However, the whole benchmark is developed as an open-source project on GitHub, without apparent communication in the AI or even NLP community. Consequently, the benchmark does not seem to trigger much emulation.

Limitations and uncovered tasks from AI evaluations

Uncharted tasks

AI and robotics are heralded as transformative, but understanding their professional limitations is crucial. A fundamental restriction is AI's reliance on function optimisation; any problem needs a clearly defined function to be tackled. Given that real-world problems often resist such simplification, AI has limited applicability in various domains, including the labour market. AI's reliance on data adds another challenge. Acquiring vast and accurate datasets is difficult, and AI systems trained on these datasets can inherit their constraints.

For instance, the O*NET Data Descriptors can help shed light on what AI and robots can and cannot do professionally. O*NET divides abilities into physical, psychomotor, cognitive and sensory. Some skills can be easily mapped to AI (e.g. speech recognition, vision tasks). Others, like inductive reasoning, cannot be tied to specific AI tasks. Several professional skills, including soft skills and adaptability, remain difficult for AI to replicate. AI tends to serve specific, narrow tasks rather than comprehensive roles, often aiding humans rather than replacing them.

Other AI capabilities

Despite advances, several application domains lack official benchmarks. Notable gaps include ML in manufacturing, as well as applications in agriculture, finance, health care, science and transportation (Sharp, Ak and Hedberg, 2018^[26]). As AI continues to evolve, the efficacy of simulations in training AI, especially robotics, becomes paramount. Early experiments have shown mixed results, indicating the need for continued exploration (Balakirsky et al., 2009^[27]).

Explainability, or an AI's ability to justify its decisions, is an emerging concern. The rise of deep neural networks, functioning as "black boxes", has increased the demand for AI transparency. However, the field of Explainable AI (XAI) remains nascent, lacking comprehensive benchmarks and evaluations.

Another growing concern is the environmental and societal impact of large-scale AI models. Their massive carbon footprints and potential to shift research towards privatisation raise questions about the sustainability and inclusiveness of the field. There is a budding interest in "frugal AI", focusing on models that use power and consume data efficiently. However, without substantial demand, large-scale evaluations and benchmarks for such models remain unlikely.

Conclusion

Leading metrology institutes are developing metrics and evaluation methods to advance research and adoption of AI algorithms for a broad spectrum of applications. This paper has summarised evaluations by LNE and NIST. As many of the evaluation discussions show, such targeted measures of performance guide and foster advancement in their target technologies. It is valuable to organise the universe of evaluations in a taxonomy to identify gaps and understand the overall landscape. This paper is an initial step towards such a taxonomy.

Further work is needed to complete the documentation of evaluations. The elements of LNE and NIST evaluations are initially merged into a single framework in this document. The scope and maturity of each evaluation must also be characterised to provide a complete picture of the state of AI and robotic skills. This review shows that such evaluation campaigns provide a wealth of evaluation data that might contribute to comprehensive measures of AI capabilities.

References

- Ajili, M. et al. (2016), *FABIOLE, a speech database for forensic speaker comparison*. [11]
- Avrin, V. et al. (2020), *AI evaluation campaigns during robotics competitions: The METRICS paradigm*, Publisher, City. [12]
- Bagchi, S. et al. (2022), “Workshop Report: Novel and Emerging Test Methods and Metrics for Effective HRI, ACM/IEEE Conference on Human-Robot Interaction, 2021”, *NIST Interagency/Internal Report (NISTIR)*. [21]
- Balakirsky et al. (2009), *Advancing manufacturing research through competitions*. [27]
- Ben Jannet, M. et al. (2014), *ETER: A new metric for the evaluation of hierarchical named entity recognition*. [13]
- Bernard, G. et al. (2010), *A Question-answer Distance Measure to Investigate QA System Progress..* [14]
- Blanco, M. et al. (2016), *Automated Vehicle Crash Rate Comparison Using Naturalistic Data*, <https://doi.org/10.13140/RG.2.1.2336.1048>. [18]
- Boothroyd, G., P. Dewhurst and W. Knight (2010), *Product Design for Manufacture and Assembly*, CRC Press, <https://doi.org/10.1201/9781420089288>. [19]
- Brunessaux, S. et al. (2014), “The Maudor Project: Improving Automatic Processing of Digital Documents”, *2014 11th IAPR International Workshop on Document Analysis Systems*, <https://doi.org/10.1109/das.2014.58>. [10]
- Costa, L. (2021), “Further generalizations of the Jaccard index”, *arXiv:2110.09619*. [15]
- Cowley, H. et al. (2022), “A framework for rigorous evaluation of human performance in human and machine learning comparison studies”, *Scientific Reports*, Vol. 12/1, <https://doi.org/10.1038/s41598-022-08078-3>. [5]
- Diño, G. (2017), *3 Reasons Why Neural Machine Translation is a Breakthrough.*, <https://slator.com/3-reasons-why-neural-machine-translation-is-a-breakthrough/#8203:%60%60oacite:%7B%22number%22:1,%22metadata%22:%7B%22type%22:%22webpage%22,%22title%22:%223>. [4]
- European Commission (2021), “Proposal for a regulation of the European Parliament and of the Council: Laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts”, Vol. 2021/0106(COD), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>. [2]
- Galibert, O. et al. (2014), *The ETAPE speech processing evaluation*. [8]
- Kahn, J. et al. (2012), *A presentation of the REPERE challenge*, <https://doi.org/10.1109/CBML.2012.6269851>. [9]
- Li, Y. et al. (2016), *MEC 2016: The multimodal emotion recognition challenge of CCPR 2016*, https://doi.org/10.1007/978-981-10-3005-5_55. [24]

- Luo, J. et al. (2021), *Robust Multi-Modal Policies for Industrial Assembly via Reinforcement Learning and Demonstrations: A Large-Scale Study*, [20]
<https://doi.org/10.15607/RSS.2021.XVII.088>.
- Mariko, D. et al. (2020), “Financial Document Causality Detection Shared Task (FinCausal 2020)”, *arXiv:2012.02505*. [22]
- Marvel, J. et al. (2020), “Towards effective interface designs for collaborative HRI in manufacturing”, *ACM Transactions on Human-Robot Interaction*, Vol. 9/4, [17]
<https://doi.org/10.1145/3385009>.
- NIST (2019), “U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools”, [3]
https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf.
- Norton, A., E. Messina and H. Yanco (2021), “Advancing capabilities of industrial robots through evaluation, benchmarking, and characterization”, in *Manufacturing In The Era Of 4th Industrial Revolution: A World Scientific Reference (In 3 Volumes)*, [16]
https://doi.org/10.1142/9789811222849_0013.
- Povolný, F. et al. (2016), *Multimodal emotion recognition for AVEC 2016 challenge*, [23]
<https://doi.org/10.1145/2988257.2988268>.
- Sharp, M., R. Ak and T. Hedberg (2018), “A survey of the advancing use and development of machine learning in smart manufacturing”, *Journal of Manufacturing Systems*, Vol. 48, [26]
<https://doi.org/10.1016/j.jmsy.2018.02.004>.
- Srivastava, A. et al. (2022), “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models”, *arXiv:2206.04615*. [25]
- Strickland, E. (2019), “IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care”, *IEEE Spectrum*, Vol. 56/4, <https://doi.org/10.1109/MSPEC.2019.8678513>. [6]
- Thrush, T. et al. (2022), *Dynatask: A Framework for Creating Dynamic AI Benchmark Tasks*, [7]
<https://doi.org/10.18653/v1/2022.acl-demo.17>.
- Van Roy, V. (2020), *AI watch-national strategies on Artificial Intelligence: A European perspective in 2019*. [1]

Annex 7.A. Low functionality levels AI tasks of evaluation campaigns across the three major fields of NLP: computer vision and robotics

Annex Table 7.A.1. Low functionality level tasks of evaluation campaigns associated with the NLP field

Sub-field	Task	Nature of the input data (application area)	Evaluation campaign name (organisers)	Integration level	Difficulty
Text comprehension	Named entity recognition	Text and audio journalistic sources (modern and 19th century), tweets, articles comment section	QUAERO[1] (LNE),	End-to-end	50% < success rate < 94% (depending on the type of named entity).
			ETAPE (LNE)		
			IMM (LNE)		
			REPERE (LNE)		
		Journalistic texts	MUC		
			ACE (NIST)		
	Story segmentation	Newswires	Topic Detection and Tracking (TDT) (NIST)	Pipeline	<60% success rate.
	First story detection	Newswires	Topic Detection and Tracking (TDT) (NIST)	End-to-end	~85% success rate.
	Story linking	Newswires	Topic Detection and Tracking (TDT) (NIST)	Pipeline	~85% success rate.
	Extraction of relations between textual phrases	Administrative documents	MAURDOR (LNE)	End-to-end (rule-based system)	~60% success rate.
Speech Processing	Diarisation	Audio debate	ALLIES (LNE),	Pipeline	~75% success rate depending on the type of input. Some systems, on some input, can go as high as 95%, but it is not the norm.
			QUAERO (LNE),		
			REPERE (LNE),		
			ETAPE (LNE)		
		Audio broadcast news, conversational telephone speech, meeting room speech	Rich Transcription (NIST)		
		Forensics, conversational telephone speech	Speaker Recognition (NIST)		
	Language identification	Administrative documents	MAURDOR (LNE)	End-to-end	~90% success rate, depending on the languages considered.
Conversational telephone speech		Language Recognition (NIST)			
	Acoustic events recognition	Audio debate	ETAPE (LNE)	End-to-end	~70% +/- 10% depending on the input (noise level, types of event).
	Story segmentation	Audio broadcast news	Topic Detection and Tracking (NIST)	Pipeline	~70% success rate on dialogues (as opposed to monologue).

Sub-field	Task	Nature of the input data (application area)	Evaluation campaign name (organisers)	Integration level	Difficulty
	First story detection	Audio broadcast news	Topic and Detection (NIST) Tracking	End-to-end	~35% success rate.
	Story linking	Audio broadcast news	Topic and Detection (NIST) Tracking	Pipeline	~80% success rate.

Annex Table 7.A.2. Low functionality level tasks of evaluation campaigns associated with the Computer Vision field

Sub-field	Task	Nature of the input data (application area)	Evaluation campaign name (organisers)	Integration level	Difficulty
Recognition	Image classification	Administrative documents.	MAURDOR (LNE), QUAERO (LNE)	End-to-end	75% < success rate < 99% depending on the classes and noise level.
	Shape recognition	Images of underwater infrastructures taken from the operating robot (underwater inspection and maintenance robotics).	RAMI (LNE)	End-to-end	~75% success rate, depending on the metrics used.
	Pose estimation	Images of industrial parts taken from a fixed angle in the workbench (agile production robotics).	ADAPT (LNE), E2919-14 Standard Test Method for Evaluating the Performance of Systems that Measure Static, Six Degrees of Freedom (6DOF), Pose (NIST)	End-to-end	40% < success rate < 99% depending on the type of input.
Motion Analysis	Shape recognition	Underwater robot camera feed (underwater inspection and maintenance robotics).	RAMI (LNE)	End-to-end	~85% success rate.

Annex Table 7.A.3. Low functionality level tasks of evaluation campaigns associated with the Robotics field

Sub-field	Task	Nature of the input data (application area)	Evaluation campaign name (organisers)	Integration level	Difficulty
	Balancing	Robot keeping balance on a vibrating plate at different vibration frequencies and amplitudes. The energy expended is measured for each round. Tests were performed in a climatic chamber through a range of temperatures.	ROBOCOM++ (LNE)	End-to-end	100% success rate, energy expenditure doubled.
	Swimming	Underwater swimming in a shallow seawater basin of 50x50m.	RAMI (LNE)	Pipeline	Good performance.
	Navigation	Aerial navigation. Performance is measured with a mean square error on position and orientation.	RAMI (LNE)	Pipeline	Good performance.
		With GNSS in agricultural field (agricultural robotics).	ACRE (LNE)	Pipeline	Good success rate but slow.
	Searching areas	Searching maze as a Man-Machine team (assuming teleoperational control).	E2853-12(2021) Standard Test Method for Evaluating Emergency Response Robot Capabilities: Human-System Interaction: Search Tasks: Random Mazes with Complex Terrain (NIST)		Low to medium success with full autonomy.
Manipulation	Pick-and-place	Pick a pole from a console and place in in a different location (underwater inspection and maintenance robotics).	RAMI (LNE)	Pipeline	50-80% success rate, strongly impacted by lighting conditions.
	Task-oriented grasping	Grasping predetermined objects and giving them an imposed position and orientation.	HEART-MET (LNE)	Pipeline	Up to 90% success rate but dependent on the computer vision algorithm.
		Robots have to autonomously assemble a defined kit of parts following a defined procedure.	ARIAC Benchmark Scenario 1: Baseline Kit Building (NIST), ARIAC Benchmark Scenario 2: Dropped parts (NIST), ARIAC Benchmark Scenario 3: In-process kit change (NIST), ADAPT (LNE)	Pipeline	No results from ADAPT physical campaigns yet.
	Hand an object over	Object placed in the robot's gripper, with the robot being placed in front of a person (assistive robotics).	HEART-MET (LNE)	End-to-end	Good success rate but poor performance (unnatural handover, slow).
	Receive an object	Robot placed in front of a person who is holding an object (assistive robotics).	HEART-MET (LNE)	End-to-end	Good success rate but poor performance (unnatural handover, slow).
	Pouring	Pouring a fluid from one container into another (assistive robotics).	HEART-MET (LNE)	Pipeline	100% success rate for a known container (End-to-end programmed motion), much lower for variable containers.
	Maintaining contact	Stay in touch with a pipe despite environment disturbances (underwater inspection and maintenance robots).	RAMI (LNE)	Pipeline	50% success rate, dependent on lighting conditions.

Annex 7.B. Detailed facet characteristics attributions of the LNE and NIST evaluations

Annex Table 7.B.1. Attribution of the facets' values to the 8 different campaigns from LNE and NIST, available online.

StatLink  <https://stat.link/w5zkxu>

Notes

¹ Usually, these adjustments tend to simplify the task from the complexity of the real world condition. Thus, they create a gap between the set-up in which systems operate compared with humans. There is no set methodology to assess the difficulty of a task and the challenge of a new evaluation. However, it has been empirically observed to be more efficient to start from the state of the art and devise a smooth progression curve. This process adds up related tasks incrementally rather than initiating short-term campaigns on disjointed, disruptive tasks. As a consequence, the capabilities of AI systems are tightly bound to the datasets available (including their annotation schemas), the physical setting and test artefacts, and the evaluation protocols defined.

² Available at <https://www.nist.gov/itl/iad/mig/metrics-machine-translation-evaluation> (accessed on 24 October 2023).

³ Available at https://www.astm.org/E2826_E2826M-20.html (accessed on 24 October 2023).

⁴ Available at <https://www.nist.gov/el/intelligent-systems-division-73500/ieee-sg-metrology-human-robot-interaction> (accessed on 24 October 2023).

8

Towards a synthesis of language capability in humans and AI

Yvette Graham, Trinity College Dublin

Edited by: Nóra Révai, OECD

Language is a major part of human intelligence and researchers have been focusing strongly on developing such competences of machines. Natural Language Processing (NLP) technologies are a key area of artificial intelligence (AI). This chapter develops a conceptual framework of language competence that allows for comparing human and machine competences. It then maps major available language benchmarks on the framework and discusses the performance of state-of-the-art AI systems on a range of tasks in two broad domains: language understanding and language generation. This exercise is a first step in building an index of AI language capability.

Language ability in humans is a major part of intelligence. Throughout human history, and far earlier than the invention of the computer, people have fantasised about building robots that can communicate and understand language. The eventual success of computers to communicate flawlessly through natural language will greatly impact society and how people work. The field of Natural Language Processing (NLP) is concerned with allowing computers to process and simulate an understanding of natural language in spoken or written form. NLP thus forms a major component of artificial intelligence (AI), but itself comprises several different sub-areas that separate NLP into problems or *tasks*. Each of these areas relates to some notion of human language competence. This raises the question: how can human language competence (which itself varies from one individual to another) be compared with state-of-the-art performance of NLP systems?

This chapter compares human levels of competence with NLP system performance levels using benchmarks from the field of computer science. In terms of the range of human language competence levels, the analysis concerns the mainstream working population and education of the general future workforce (i.e. it reflects the competences of any human who does not possess a severe disability).

Benchmark tasks: Narrow versus strong AI

Each research area in NLP includes one or more *benchmarks* or *shared tasks* that aim to compare the performance of competing approaches to determine which methods have most promise. A task is defined with a core research goal of automating some form of language processing in a specific way and with respect to a specific domain of language (e.g. translating news documents from German to English). To correctly interpret NLP benchmark test results, *narrow AI* that focuses on solving individual tasks must be distinguished from *strong AI* that aims to simulate *general purpose intelligence*. Technologies are developed and tested in isolation from other tasks so almost all NLP benchmarks currently form part of narrow AI. This means that NLP systems can legitimately be tested on their performance of a single language task within a single domain (e.g. news text, medical documents, literature or scientific papers). The performance achieved – even if very high – is limited to the evaluation setting, which is restricted to that specific domain and task. Working on individual problems in NLP in isolation from other tasks allows for progress, making an insurmountable mountain climbable through mostly independent routes.

While it is important to interpret NLP success within the context of narrow AI, components of a successful NLP system can often be applied to a new task, domain or language. Such components may very well form part of an eventual general purpose language AI. How the research community can ever achieve a general purpose (or strong) language AI is still a question. The key to this may lie in the underlying technologies that have proven successful (or will be) across multiple NLP tasks, languages and domains (see recent development in Box 8.1).

Box 8.1. Transformer models

The recent development of neural NLP architectures as transformer models has helped the move towards general purpose AI. The emergence of this paradigm has been a game changer in terms of NLP system performance, resulting in discussions around human parity at several tasks. A transformer model learns to understand and represent the meaning of language from an exceptionally large volume of raw text with no human annotation.

The most well-known such model developed by OpenAI, GPT-3 (and its first publicly released version, ChatGPT) is trained on half a trillion words (or tokens) of English, sourced from a combination of webpage content and books. This massive language model requires training only a single time and can then be applied repeatedly and in a wide range of distinct tasks. The model can be deployed to successfully automate a range of distinct language tasks, such as Machine Translation, Named Entity Recognition, Question Answering and Speech Recognition, among others.

Typically, a single pretrained language model can be used with further fine-tuning, which requires only a relatively small data set. This model is referred to as a *transformer* since it transforms the information it has learnt from the pretrained model to a more specific task. The technology behind this approach is based on the structure of the human brain in the form of neural networks.

Conceptual framework of language competences

Comparing human language competence to NLP benchmarks requires bridging the gap between the general understanding of description and analysis of human language competence and NLP research.

Annex Table 8.A.1 presents a list of main NLP research areas with the equivalent human skill each task aims to automate. Human language competence is generally not analysed by tasks. Rather, it is usually divided into the four competence areas of reading (HR), writing (HW), listening (HL) and speaking (HS). Language competence is then described on a range from low level of competence in a given native language (e.g. with no reading or writing ability) to high proficiency in all four categories in the native (or another foreign) language.

To map human language competences to NLP areas, two high-level groups can be formed:

- *Language understanding*: NLP tasks that correspond closely to reading or listening competence in humans (understanding/comprehension tasks); both require interpretation of *input*.
- *Language generation*: NLP tasks that correspond to writing or speaking require the system to *output* language.

Some NLP tasks correspond to a combination of language understanding and generation. The distinction between reading or listening and writing or speaking is only a matter of input/output formats moving from text to speech. The core technology or research problem, such as translation, largely remains the same regardless of input or output format.

However, the format that an NLP system receives or produces can impact system performance. A move from text input/output to spoken input/output usually involves a *decrease* in system performance, as spoken input is less predictable and more difficult to process than textual input. The opposite is true for humans: understanding or generating spoken language instead of reading or producing text is usually easier because it does not require competence in literacy. Figure 8.1 shows this relationship. There are exceptions to this general rule. For example, machine translation and interpreting are more challenging for both system and human translation in spoken form.

Figure 8.1. General relationship between human language and NLP difficulty levels with respect to the input and output format moving from text to speech and vice versa

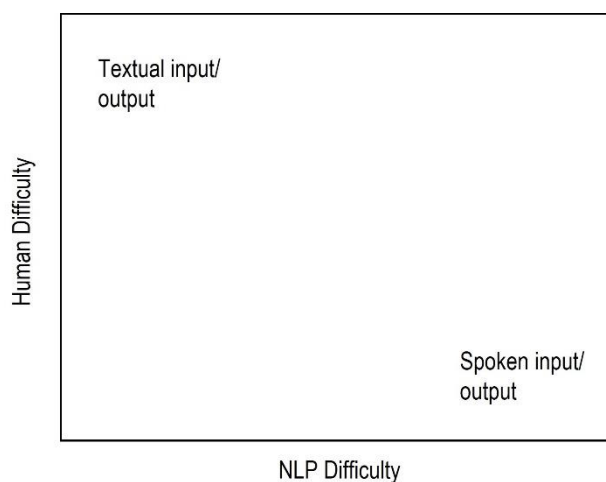


Table 8.1 provides a rough guide of how core NLP research areas relate to the type and competence level required of a human to perform the equivalent task. Some tasks require different abilities from systems and humans. For example, Speech Recognition for humans corresponds to the task of understanding spoken language and, as such, generally requires low level of language ability. However, NLP systems are tested by the production of a transcription, i.e. a written output.

Table 8.1. NLP research areas with type and level of language competence required for humans

Minimum Human Competence Level	Reading/Listening	Both Reading/Listening and Writing/Speaking
Below Average	Emotion-cause Pair Extraction Event Extraction Humour Detection Reasoning (basic) Visual QA	Dialogue (open domain) Speech Recognition
Average-High	Anaphora Resolution Natural Language Inference Part of Speech Tagging Reasoning (advanced) Sentiment Analysis Text Classification Topic Modelling	Explanation Generation Grammar Correction Keyword Extraction Lexical Normalisation Punctuation Restoration Question Answering Reading Comprehension Sentence Compression Summarisation Text Diacritisation
Specialist	Authorship Verification Automated Essay Scoring Information Retrieval Semantic Parsing Syntactic Parsing Relation Extraction	Dialogue (task-oriented) Machine Translation Text Style Transfer

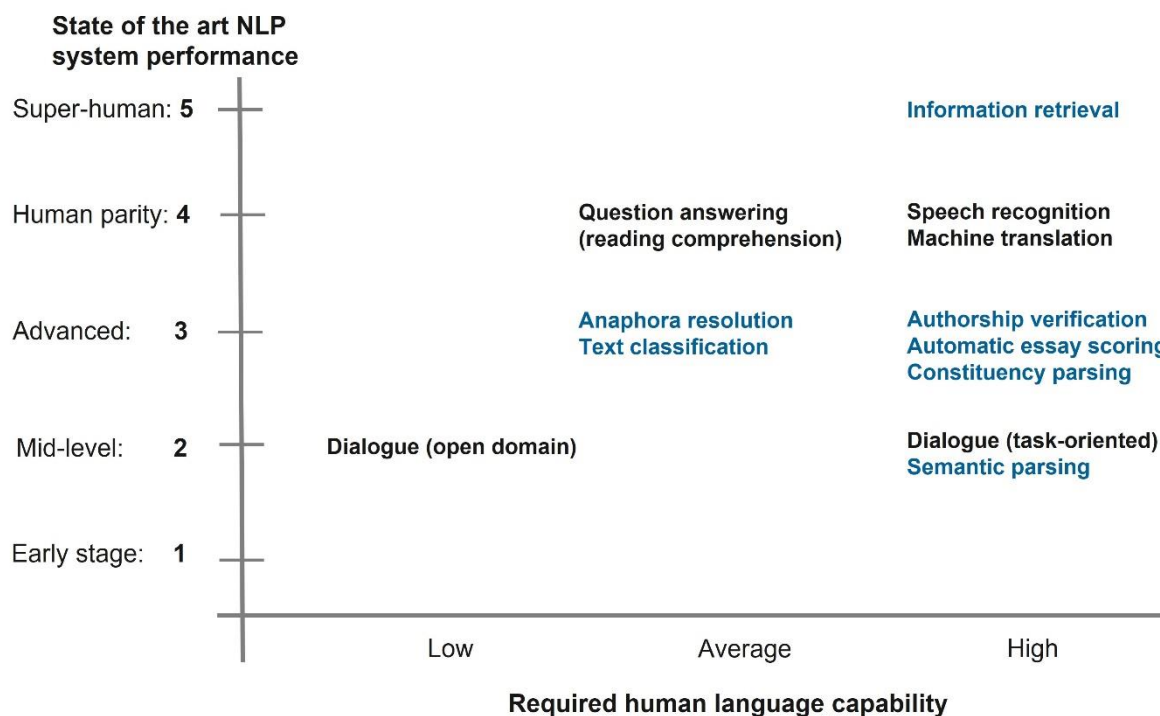
Mapping major language benchmarks to the human language competence framework

This section maps the performance level of state-of-the-art systems in each research area to human competence levels required for a corresponding task. Figure 8.2 comprises three basic performance levels corresponding to: i) *early-stage research*; ii) *mid-level research*; and iii) to *advanced research*. There are two additional categories: iv) tasks in which *human parity* discussions have begun and v) areas with consensus that systems have already achieved *super-human* performance. The figure reflects state-of-the-art research in 2021; the performance of systems is likely to improve over time.

Some NLP tasks, such as Machine Translation (MT) or Speech Recognition, are application-driven, i.e. they correspond to human tasks. For these, it is natural to ask: *how does state-of-the-art AI compare to the ability of a human completing the same task?* With these tasks, researchers aim to reach performance equivalent to that of humans (human parity), or even surpassing the capability of all humans (super-human performance).

However, not all NLP tasks are driven by direct application. Some tasks aim to automate an annotation process that usually requires a human to complete. The ultimate aim is simply to annotate/label data correctly as a human would have. The human annotator in this case provides the gold standard annotations against which the outputs of the NLP technology are judged (as either correct or incorrect). For such tasks, questions around human-parity or super-human performance are arguably not highly relevant. Therefore, these research areas are only classified in the first three categories.

Figure 8.2. Relationship between required minimum human language competence level and state-of-the-art NLP system performance for a sample of NLP tasks



Note: This is a rough guide based on state-of-the-art NLP research in July 2022. Blue benchmarks refer to language understanding, black ones to language generation.

Figure 8.2 provides a rough guide of the current relationship between NLP state-of-the-art performance and the minimum human language competence required to complete that task, with reference to a sample of NLP tasks. The sections that follow describe each task shown in the graph and explain why it was placed at the given performance level based on recent NLP benchmarks. A discussion of language understanding is followed by language generation tasks.

Language understanding: AI vs. human

Overall, Figure 8.2 suggests that AI systems perform better in language understanding tasks than a human with low level reading and listening competence. Many research areas are in an advanced stage, with Information Retrieval (IR) being at the top – super-human – level.

Understanding words in their context

Task definition: Anaphora and Coreference resolution

The meaning of many words cannot be interpreted correctly isolated from their context. For example, *it* commonly refers to a noun mentioned earlier in the text. To translate the word *it* from English to German, the gender of the word *it* refers to in the text must be known. For example, in “*The car was brand new, but he still allowed me to drive it*”, the gender of *car* in German must be known. Understanding this link between words in a text/speech is relatively easy for humans but can be challenging for NLP systems.

Anaphoric words are expressions like *it* in the example above, whose interpretation depends on the context. Anaphora and coreference resolution are NLP tasks that aim to identify entities referred to within text or spoken language by such anaphoric words or expressions.

System performance on benchmarks

The domain of anaphora and coreference resolution includes an extensive range of entities and numerous benchmarks that test system performance (see Sukthanker et al. (2020_[1])). A main dataset is the Ontonotes corpus (Weischedel et al., 2013_[2]). This was created to develop methods of automatic coreference in order to link all the specific mentions in a text that refer to the same entity or event. In addition, the corpus was annotated to distinguish between different types of coreference. Texts automatically annotated in this way are likely to help other NLP tasks learn to correctly process multiple mentions of the same entity.

State-of-the-art systems based on a transformer architecture achieve high performance with respect to Ontonotes: an F-score (a score that combines precision and sensitivity) of approximately 81% (Dobrovolskii, 2021_[3]). Despite a lack of human performance estimates for the task, performance can still be gauged to some degree. The highest performing systems surpass the performance of someone with low reading and listening competence. However, they are unlikely to surpass that of a human with an average or high language competence.

Box 8.2. Example of anaphora and coreference resolution

In the example dialogue below taken from the TRAINS corpus (Poesio et al., 2016₍₄₎), the personal pronoun *it* refers to two distinct objects in utterances 3.1 and 5.4.

Figure 8.3. Example dialogue from the TRAINS corpus

1.1 → M: all right system
 1.2 : we've got a more complicated problem
 1.4 : first thing _l'd_ like you to do
 1.5 : is send engine E2 off with a boxcar to Corning to pick up oranges
 1.6 : uh as soon as possible

2.1 → S: okay

3.1 → M: and while it's there it should pick up the tanker

4.1 → S: okay
 4.2 : and that can get
 4.3 : we can get that done by three

Source: Adapted from Poesio et al. (2016₍₄₎): *Anaphora Resolution*, Springer.

Structuring and organising text into categories

Task Definition: Text Classification

Text classification systems aim to automatically categorise a given sentence or document into an appropriate category. They can help organise and structure any kind of text, such as sentences, documents and files. The number and types of categories depend on the application and associated dataset. For example, news articles can be categorised by topic, sentences can be labelled with the emotions expressed by them or as grammatical or ungrammatical. Systems can then be trained on such datasets to classify unseen sentences in these categories. As such, text classification overlaps with other more specific NLP research areas, such as sentiment analysis and grammar correction.

System performance on benchmarks

Text classification is a long-established area of NLP with extensive work and progress likely due to the wide availability of substantial data for training and testing systems. One of the most widely used set of datasets is the General Language Understanding Evaluation (GLUE) (see Box 8.3).

Box 8.3. GLUE benchmark

The General Language Understanding Evaluation (GLUE) benchmark comprises nine datasets for classification of sentence of English. The name of each dataset and the task associated in Table 8.2 is adapted from Wang et al. (2018_[5]).

Table 8.2. Datasets in the GLUE benchmark

Name	Description
CoLA (Corpus of Linguistic Acceptability)	Determine if a sentence is grammatically correct or not.
MNLI (Multi-Genre Natural Language Inference)	Determine if a sentence entails, contradicts or is unrelated to a given hypothesis.
MRPC (Microsoft Research Paraphrase Corpus)	Determine if two sentences are paraphrases from one another or not.
QNLI (Question-Answering Natural Language Inference)	Determine if the answer to a question is in the second sentence or not.
QQP (Quora Question Pairs2)	Determine if two questions are semantically equivalent or not.
RTE (Recognising Textual Entailment)	Determine if a sentence entails a given hypothesis or not.
SST-2 (Stanford Sentiment Treebank)	Determine if the sentence has a positive or negative sentiment.
STS-B (Semantic Textual Similarity Benchmark)	Determine the similarity of two sentences with a score from 1 to 5.
WNLI (Winograd Natural Language Inference)	Determine if a sentence with an anonymous pronoun and a sentence with this pronoun replaced are entailed or not.

Source: Wang et al. (2018_[5]): *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*. <https://doi.org/10.18653/v1/w18-5446>.

State-of-the-art results can vary depending on datasets but range from approximately 75% to 90% in terms of accuracy. Despite high performance, discussions of human parity at this task have not yet taken place. This is likely because the task is closer to an annotation (labelling) task than an application task. In other words, human annotations for this task are deemed correct, leaving aside disagreement between annotators that can occur. This area can be classified at an *advanced* performance level, with at least average human language competence required to perform the same task.

Understanding the meaning and role of expressions

Task Definition: Semantic Parsing

Understanding the meaning of expressions within and across sentences can be the basis for translation, answering questions and reasoning. Semantic parsing aims to automatically annotate sentences of a given natural language with *formal meaning representations*. A typical example is to automatically identify the subject, object and indirect object in a sentence of English. For example, the sentences “*Mary gave John the letter*” and “*Mary gave the letter to John*” are identical in terms of semantic structure, despite the order of words being different.

System performance on benchmarks

Semantic parsing is an easy task for humans, even someone with low literacy skills can figure out who did what to whom in the sentence above. However, the problem is more challenging for machines. They require parsing the sentence with the grammar of the specific language to determine this simple information from a simple sentence. There are numerous possible formalisms for semantic parsing of natural language, and studying a single natural language, such as English, only illustrates a fraction of the features of

language in general. Further challenges include long-distance dependencies between words (i.e. links between words that are far from each other in the sentence) and languages that suffer from data sparseness due to rich morphology, such as Arabic, Czech and Turkish. Interestingly, the original applications of semantic parsing, such as MT, have had much more success without the integration of semantic representations.

Parsing language has received a good amount of attention over the years within the NLP community. A number of datasets for training and testing semantic parsers exist for a range of different formalisms. Propbank (Palmer, Gildea and Kingsbury, 2005^[7]) is a corpus annotated with predicate argument structure. For its part, Framenet is a major dataset (Baker, Fillmore and Lowe, 1998^[8]) in which the usual unit of meaning – a word – is replaced with other lexical units and frames. More recently, a project known as Universal Dependencies has made gallant strides towards developing a cross-linguistically consistent treebank annotated with semantic roles with the goal of multilingual parser development (de Marneffe et al., 2021^[9]).

Although extensive time and energy have been invested in this research area and high accuracy achieved in shared tasks, much work to date has focused on repeated tests on the same dataset. In addition, tasks have overly focused on English, a language that probably poses far fewer challenges than morphologically rich languages. As a result, correctly classifying semantic parsing technologies is difficult. This area could be best placed as having *mid-level* performance to consider the remaining challenges for developing technologies to a large number of languages. Corresponding minimum human level of language competence is high (specialist) in terms of formal annotation (and not simply understanding a sentence).

Understanding sentence structure

Task Definition: Constituency Parsing

Analysing a sentence by breaking it down into sub-phrases of different grammatical categories (e.g. noun phrases, verb phrases) can help in more complex language tasks, such as grammar checking, semantic analysis and Question Answering (QA) (Jurafsky and Martin, 2023^[10]). Constituency parsing is the task of automatically annotating sentences of natural language with a phrase structure grammar. Figure 8.4 shows a constituency parse tree corresponding to a phrase structure grammar of an example sentence.

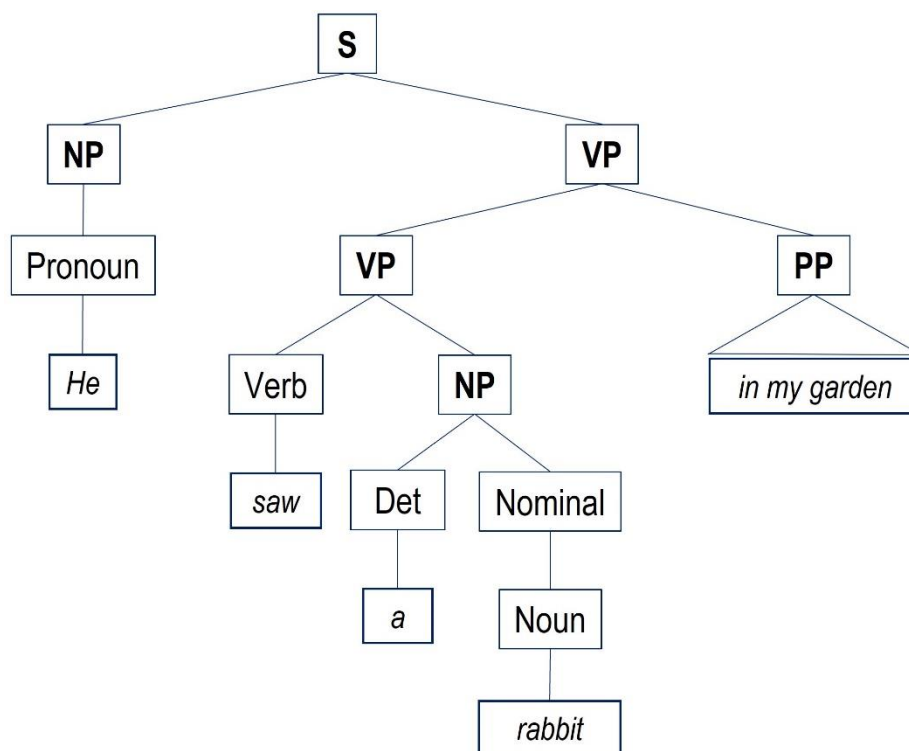
System performance on benchmarks

Two major datasets for testing constituency-parsing technologies include the English Penn Treebank and the Chinese Treebank, with highest performing systems achieving approximately 97% and 93%, respectively (Mrini et al., 2020^[11]).

Since constituency parsing is essentially automated annotation of data (annotation task), the question of human parity is unlikely to be discussed. The question is not whether systems have reached human parity in constituency parsing itself. It is more interesting to ask whether systems have helped reach human parity with respect to a larger application task.

This research area falls into *advanced* performance level with respect to parsing English text, since accuracy is exceptionally high. However, this research area has shown an excessive focus on parsing results for English.

Figure 8.4. Sample constituency parse tree of English sentence



Source: Adapted from Jurafsky and Martin (2023_[10]): *Speech and Language Processing*, Pearson Education Inc.

Assessing student essays

Task Definition: Automatic Essay Scoring

Automatic Essay Scoring (AES) is the task of automatically assessing student essays and has applications within education. For example, AES has great potential to allow students to get feedback about the quality of their writing without requiring teaching professionals' time and resources.

System performance on benchmarks

The Automated Student Assessment Prize dataset, released by Kaggle (www.kaggle.com), is a main resource for training and testing AES systems (www.kaggle.com/competitions/asap-aes/data). The dataset contains eight essay sets, with essays ranging from 150 to 550 words per response, written by students in US grade levels from grades 7 to 10. Essays within the dataset are hand-graded and double-scored to test rater reliability.

Results for state-of-the-art systems show performance at approximately 80% weighted kappa. Discussions of human parity in this area are unlikely to develop. Arguably, they are not highly relevant due to the nature of automating a task for which human scores provided by a qualified teacher are considered valid and correct. AES can be considered as having *advanced* system performance while requiring high (specialist) human language competence.

Identifying the author of a document

Task Definition: Authorship Verification

Authorship verification (AV) is an NLP task to automatically determine if a new unseen document was authored by an individual already known to the system. The task has applications in detecting plagiarism and in data analytics for commercial systems that aim to profile users based on the content they have authored on line, among others.

System performance on benchmarks

Despite the huge potential of NLP technologies to successfully categorise textual content according to author, testing in this area has been limited by lack of data. Data needed for this task would ideally comprise text authored by a large set of individuals (n), each of whom had authored a large set of documents (m), yielding a potential dataset containing $n \times m$ texts to train systems. Such datasets are unfortunately not readily available. Furthermore, even given the availability of such data for a single domain, systems should be tested on texts/documents across a range of domains of language to verify that results achieved for one domain can be achieved in another.

To work around the lack of ideal training and test data, benchmarks have taken the available data and simplified the task. They only require systems to determine if the same individual authored two given documents (Göeau et al., 2021^[12]). Data for benchmarks are taken from the fanfiction domain. Fanfiction refers to new stories authored by fans of a well-known show/book that include its characters (Kustritz, 2015^[13]). Since these kinds of data provide multiple stories authored by the same individual, fanfiction lends itself to training and testing AV systems.

Despite systems generally only being tested within the fanfiction domain, there is no reason to believe that systems would not achieve similar results in other domains provided that data are available. However, measures applied in benchmarks for this task are not straightforward to interpret, nor do they readily map to human competence. For example, systems are permitted (and somewhat encouraged) to sit on the fence for test items and submit decisions of 0.5 probability to indicate indecision about a difficult case. This results in metrics reported on distinct numbers of outputs. This, in turn, gives an advantage to systems that sit on the fence more often, making comparisons across systems difficult. Consequently, it is difficult to gain a simple intuition about how often systems correctly identify authors of documents.

In addition, no attempts have been made to estimate the degree to which humans can determine if two stories had the same author. This leaves the question about the performance level of state-of-the-art AV systems. The organisers of the latest shared task report that the F-score of the best system (about 95%) is highly encouraging. However, they note that these results may not hold for other domains and the test domain may simply be too easy.

This task was placed as having *advanced* system performance, while keeping in mind the above-discussed caveats. Even humans with high language competence might find this task exceptionally challenging. If this is the case, systems could perform better than humans at AV.

Finding material relevant to a question

Task definition: Information Retrieval

Information Retrieval (IR) is defined as finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) (Manning, Raghavan and Schütze, 2008^[14]). IR is not generally considered to be a sub-discipline of NLP, as the methods proven successful especially in the early days have had relatively little in common with

NLP systems. IR has focused on word counts in documents (or statistics), compression techniques and efficient algorithms to speedily sift through enormous quantities of data to respond to the information need of a user.

System performance compared to humans

Even the most basic IR algorithms from the early days have already surpassed human limits for IR due to the scale of data needed to be searched. Benchmarks are thus less relevant to understand the extent to which IR has achieved performance comparable with humans. IR systems clearly have super-human performance. A human carrying out the IR task corresponds to the task traditionally carried out by a trained librarian, and thus requires high (specialist) language competence.

Language generation: AI vs. Human

The competence to generate language whether in spoken or written form generally requires language understanding in real life and workplace language tasks. Figure 8.2 suggests that AI performance in language generation ranges from mid-level to human parity depending on the research area.

Dialogue in an open domain

Task Definition: Open-Domain Dialogue

Dialogue systems aim to automate the art of human conversation. Dialogue research is generally split into two distinct areas. Open-domain dialogue automates conversation about any topic of interest, also referred to as chit-chat models. Task-oriented dialogue completes specific tasks through automated conversation (see section Dialogue in a narrow domain).

System performance on benchmarks

A main venue for evaluating dialogue systems is the Conversational Intelligence Challenge (Convai) (Burtsev et al., 2018_[15]; Dinan et al., 2019_[16]). Evaluating open-domain dialogue systems is challenging because it requires human evaluators to talk to chatbots about an open or prescribed topic before rating their quality. In fact, human evaluations in the original competition were found to be unreliable, and the results discussed here use an alternate source for evaluation results.

A recent evaluation of state-of-the-art models, which was replicated to ensure reliability, showed that the highest performing models rated by human judges perform at approximately 52% (Ji et al., 2022_[17]).¹ State-of-the-art models in this area are based on transformers that learn from extremely large language models. They produce highly fluent output that often is appropriate for the input provided by its human conversation partner. However, despite fluent output, models still lack consistency in conversations and the ability to reliably incorporate knowledge or learn new information from conversations. We place this research area into the category of *mid-level* performance in relation to other NLP tasks. This task needs low language competence from humans.

Understanding and transcribing spoken language

Task Definition: Speech Recognition

Speech recognition systems aim to take in a speech signal from one or more speakers and produce a textual transcription of the language that was spoken. Much of the research developed in speech

recognition has been applied successfully to other NLP tasks. As mentioned previously, systems are tested on their transcription ability, while testing humans' speech recognition does not require written literacy. This complicates the comparison, as for machines it involves the generation of language output.

System performance on benchmarks

Speech recognition has received a wealth of attention over the years, partly due to the availability of data for training and testing statistical systems. There are a large number of benchmarks and datasets for this area and it is a leading area of NLP in terms of system performance. Indeed, many other research areas adapt methods first successful in speech recognition. Researchers are discussing human-parity and even super-human performance of systems, but challenges nonetheless remain. It is yet to be shown that the accuracy of developed speech recognition technologies exceeds that of a human transcription for all kinds of spoken language data. Speech recognition was placed at the human-parity level. Since almost all humans (without serious disabilities) have this capability, it only requires a low human language ability.

Answering questions based on reading comprehension

Task Definition: Question Answering

Question answering (QA) is the task of automatically finding answers to questions or identifying when a question cannot be answered due to ambiguity or lack of information to provide an appropriate answer. QA overlaps with other NLP research areas such as reading comprehension. Rather than mere IR from a pre-engineered knowledge base with facts, QA with reading comprehension means a system can comprehend a text and absorb the knowledge without prior human curation.

System performance on benchmarks

QA is an extensively studied area of NLP. There is a vast number of datasets and benchmarks for evaluating systems, likely exceeding 100 distinct datasets for testing some form of QA system. The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016_[18]), for example, is a reading comprehension data set containing a sample of questions and answers about Wikipedia articles. In SQuAD, systems must either find the answer to each question in the corresponding article or identify that the question is, in fact, unanswerable. Results for state-of-the-art QA systems for SQuAD and other QA datasets such as CommonsenseQA reveal excellent system performance and have led to discussions of systems reaching human parity. The human language competence required for effective QA is average to high.

Dialogue in a narrow domain

Task Definition: Task-oriented Dialogue

Certain conversations pertaining to a specific task, such as asking about the weather and giving navigation instructions to a driver, happen regularly in many people's lives and are thus worth automating. Task-oriented dialogue systems aim to help users complete a task of some description through automated conversation. This can be in the form of actual spoken interaction with the system or a text-based interface or task-oriented chatbot.

System performance on benchmarks

Given the large number of use cases, testing for task-oriented dialogue systems is still limited by available training and test datasets. A main dataset for task-oriented dialogue is Key-Value Retrieval Networks (KVRET), which contains more than 3 000 dialogues across the in-car assistant domain, calendar

scheduling, weather IR and point of interest navigation in English (Eric et al., 2017_[19]). Models are evaluated using entity-F1, a metric that evaluates the model's ability to generate relevant entities from an underlying knowledge base and to capture the semantics (Eric et al., 2017_[19]). The current best task-oriented dialogue system achieves an entity F1 score of approximately 71% (Xie et al., 2022_[20]).

Humans have achieved 75% on such tasks, which suggests impressive performance of state-of-the-art systems. However, AI systems were evaluated on datasets similar to training data, instead of new unseen test data. Given the limitations of evaluations and available datasets system performance must be judged with caution. Task-oriented dialogue is still extremely challenging, corresponding most closely to *mid-level* performance. On the human side, such tasks often require a specialist and thus correspond to high-level language competence.

Translating text

Task Definition: Machine Translation

Machine Translation (MT), i.e. automatically transferring the meaning of text or speech from one natural language into another, is one of the earliest AI language tasks. It presents many challenges. First, there are, in theory, an infinite number of possible input sentences. This means that translating sentences requires breaking them down into short units to find a *phrase* that has been seen in the training data (or is in the database), and translate it by components (slicing and dicing). Second, there are many ways to slice and dice a single sentence, and many possible ways to translate each of those slices. This results in a vast number of possible outputs for every input sentence. Determining which of these is the best translation is a main challenge in MT.

While phrase-based MT has been a highly successful and long-standing approach to translation (Koehn, Och and Marcu, 2003_[21]), transformer models have recently made substantial advances. Despite this progress, challenges remain. These include sentences containing long-distance dependencies between words, issues relating to incorrectly translated pronouns (discussed in the section on anaphora resolution), the translation of languages with rich morphology and languages with low amounts of training data.

System performance on benchmarks

Generally in MT, the two languages between which a system translates is known as its *language pair*. Besides the many challenges that lie within the task of MT, language pairs create another problem with respect to reporting system performance. MT performance can be easily gauged from benchmark results for a language pair that, for example, translates from German (de) to English (en). However, assessing translation performance between *any* two natural languages is less straightforward. This is partly because of the large number of possible language pairs, but more importantly because of big performance gaps between language pairs. MT has excellent performance for some pairs, for which large datasets are available and researchers have worked on them extensively. Conversely, performance is much lower for pairs that have no data or systems for testing.

The data-driven methods the MT research community has most focused on are language pair independent. These state-of-the-art methods learn to translate from large corpora (as opposed to hand-crafting large sets of rules). This means that once training data for many language pairs become available, a high performing system can be built relatively quickly using already developed code. Therefore, with some degree of caution, results can be extrapolated on state-of-the-art methods for which training data already exist for any language pair.

Recent advances have resulted in discussions about human parity in many leading benchmarks. The news translation task at the Conference on Machine Translation (WMT) (Akhbardeh et al., 2021_[22]), for example, has highly valid and reliable test results. This is because substantial effort is put in developing new test

data before each annual competition, ensuring that the test data are truly unseen. In addition, the task employs human evaluation as opposed to automatic scoring, with both systems and human translators included in competitions in a blind test. WMT benchmarks have shown that on average the best system(s) achieve performance on-par with a human translator.

Update of AI language competences post ChatGPT release

The first draft of this chapter was written in summer 2022. In November 2022, ChatGPT was released, marking a significant milestone in AI. By January 2023, it garnered over 28 million daily visits, which many describe as the highest impact AI technology advancement. OpenAI presents ChatGPT as a step towards Artificial General Intelligence. It can hold high-quality conversations, answer follow-up questions, admit mistakes and challenge incorrect assumptions. Many users find its simulated intelligence convincing, able to answer any question with higher fluency and linguistic ability than many speakers of English. ChatGPT has been tried for tasks like writing job applications, emails and even academic essays.

Despite the apparent advancement in recent dialogue systems, such as ChatGPT, **it is not possible to track significant improvement in AI's language capabilities over one year**. This is because new models have not yet been independently tested on language benchmark tasks.

This brief section re-evaluates the benchmarks used to measure AI language capabilities and systems' performance on the old and new benchmarks in light of the advancements.

Evolution of benchmarks

While all benchmarks discussed above remain pertinent, some already existing and new benchmarks have gained importance. SuperGlue, an advanced version of the original GLUE benchmark, has emerged as particularly significant. It offers a comprehensive metric for gauging progress toward general-purpose language understanding for English (Wang et al., 2018^[5]). SuperGlue's new tasks provide a better measure of AI's underlying capabilities compared to its predecessor. Tasks included in SuperGlue are challenging for AI but solvable by most college-educated English speakers. SuperGlue features:

- more challenging tasks, retaining the two toughest from GLUE and adding others based on their difficulty for current NLP systems
- diverse task formats, expanding from just sentence pair classification to coreference resolution and QA formats
- comprehensive human baselines
- enhanced code support
- refined rules to ensure fair competition.

Importantly, SuperGlue provides human performance estimates. In 2019, the average accuracy across eight tasks was 71.5, compared to a human score of 89.8, indicating an almost 18 percentage point difference (Wang et al., 2018^[5]). More recently, in October 2023, the leaderboard shows significantly improved performance figures for more competitive systems, reflecting a large increase in performance since 2019. However, numbers need to be interpreted with a degree of caution as submissions to SuperGlue are permitted up to 6 times per month, so at least some degree of tuning to the test could have taken place over the past 4 years.

However, SuperGlue has some limitations. First, tasks that require domain-specific knowledge were not included. Thus, SuperGlue is not able to assess AI's capability to integrate knowledge and language competences. Yet, this is necessary to closely mimic human intelligence. Second, some human evaluation estimates rely on anonymous crowd-sourced data, which might be of lower quality than expert annotation, potentially affecting human performance estimates.

Other new benchmarks aiming to test general-purpose intelligence have also emerged this past year. One that has gained significant attention is the Beyond the Imitation Game Benchmark (BIG-bench) (Srivastava et al., 2022^[23]). BIG-bench is a collaborative benchmark with more than 200 tasks, intended to probe large language models and extrapolate their future capabilities. However, it moves beyond language processing tasks and thus is not considered in this update. Similarly, benchmarks that focus on evaluating language models in a zero-shot setting (e.g. LMentry) were not considered either. These benchmarks test a pretrained language model's general understanding of language through transfer learning as opposed to their performance on a specific task they were trained on.

To date, there are no appropriate benchmarks available to assess technologies like ChatGPT as a single system that covers a broad spectrum of NLP tasks. Future research would benefit from benchmarks designed for general-purpose language AI systems, enabling more accurate comparisons between ChatGPT, its competitors and human capabilities.

Evolution of system performance in language

A system that aims to be an Artificial General Intelligence is not designed for one specific NLP task only. Rather, it aims to simulate human language ability across various tasks. ChatGPT focuses on a subset of NLP tasks: question answering, summarisation and general language generation, including letter writing, report writing and storytelling. Importantly, it aims to master dialogue in an open domain. Its success stems from its ability to integrate multiple tasks seamlessly, offering a user-friendly interface. However, when comparing ChatGPT to individual NLP benchmarks and human language capability, it is essential to assess its performance in specific tasks.

Overall, almost all NLP tasks explored in this paper remain at the same level of performance after about a year of development. There are two tasks that need revisiting because of the breakthrough with large language models: Dialogue in a narrow domain (task-oriented dialogue) and in an open domain.

Regarding task-oriented dialogue, AI performance has improved on a limited number of tasks that can be performed to a high standard through conversational instruction by systems such as ChatGPT. However, systems still cannot perform successfully on a wide range of tasks through dialogue with users. Thus, overall performance remains at mid-level.

As a fluent conversational agent, ChatGPT and other publicly available dialogue systems based on transformers and large language models constitute a breakthrough. However, ChatGPT still faces challenges in numerous tasks, like applying knowledge learnt in conversations with users, resolving references and integrating real-time world knowledge. Thus, overall performance of AI in open-domain dialogue may not yet have moved to an advanced level.

Other breakthrough technologies beyond ChatGPT

Following the release of ChatGPT, OpenAI introduced GPT-4, a significant advancement in NLP. GPT-4 processes both image and text inputs to produce textual responses. While independent evaluations of GPT-4 are still pending, OpenAI has shared results that suggest advanced performance. GPT-4 is reported to achieve human parity on various professional and academic benchmarks in terms of factuality, steerability and ethically declining certain queries.

GPT-4 is said to pass simulated bar exams, scoring within the top 10%, a significant improvement from its predecessor GPT-3.5, which scored around the bottom 10% (OpenAI, 2023^[24]). However, it's essential to note that these are simulated tests, suggesting they might differ from real-world exam settings. In addition, the results discussed here are OpenAI's in-house tests that have not yet been independently verified and only limited details are publicly available. Despite these accomplishments, OpenAI acknowledges GPT-4's limitations, especially when compared to human performance in real-world scenarios.

Conclusion

This chapter provided a framework that captures the relationship between the language competence of humans and AI systems. It considers the variance between human language competence levels, while focusing only on best-performing NLP systems. The chapter analysed systems' language performance in 12 selected research areas. Results suggest that in four of these areas, machines are at the level of humans or higher; in an additional five, AI is at an advanced stage of research; and in the remaining three domains, AI is at mid-level.

The chapter discussed some challenges in directly comparing human and machine NLP capabilities. First, machines are usually developed for and tested on narrow tasks rather than broad capabilities. While there are general literacy tests for people, AI systems are evaluated on specific language tasks. This makes the comparison with respect to broader capabilities challenging. Second, machines do not have the same difficulties as humans and so understanding machine competences may require different sub-areas of language competence. Despite these difficulties, the chapter provided an initial comparison of human and AI language competence.

This work has revealed that the available data for training a system to automate a task is the factor with the most influence over progress within a specific NLP research area. Policy makers or other actors could also influence the development of AI language capabilities if they identify a new language task in which AI could have high positive impact (e.g. for governments, society or commerce). Consulting experts about what data would be needed to train systems and creating a large public dataset would allow the research community to develop systems.

In sum, as systems improve in performance, the impact of NLP technologies on society is likely to be significant. The pace of development can be very fast if large datasets become available and if there is sufficient market potential. This chapter is a first step in building an index that will help the wider public understand what current technology can do and how performance of AI in language corresponds to human language competence levels. Taking this work forward will involve the analysis of a wider set of benchmarks and the development of an AI language index that can be regularly updated to inform the education policy community on AI progress in the field.

References

- Akhbardeh, F. et al. (2021), *Findings of the 2021 Conference on Machine Translation (WMT21)*. [22]
- Baker, C., C. Fillmore and J. Lowe (1998), "The Berkeley FrameNet Project", *Proceedings of the 36th annual meeting on Association for Computational Linguistics -*, <https://doi.org/10.3115/980845.980860>. [8]
- Burtsev, M. et al. (2018), "The First Conversational Intelligence Challenge", in *The NIPS '17 Competition: Building Intelligent Systems, The Springer Series on Challenges in Machine Learning*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-94042-7_2. [15]
- de Marneffe, M. et al. (2021), "Universal Dependencies", *Computational Linguistics*, pp. 1-54, https://doi.org/10.1162/coli_a_00402. [9]
- Dinan, E. et al. (2019), "The Second Conversational Intelligence Challenge (ConvAI2)", in *The NeurIPS '18 Competition, The Springer Series on Challenges in Machine Learning*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-29135-8_7. [16]
- Dobrovolskii, V. (2021), "Word-Level Coreference Resolution", *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, <https://doi.org/10.18653/v1/2021.emnlp-main.605>. [3]
- Eric, M. et al. (2017), "Key-Value Retrieval Networks for Task-Oriented Dialogue", *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, <https://doi.org/10.18653/v1/w17-5506>. [19]
- Göeau, H. et al. (2021), "Overview of plantclef 2021: Cross-domain plant identification". [12]
- Ji, T. et al. (2022), "Achieving Reliable Human Assessment of Open-Domain Dialogue Systems", *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, <https://doi.org/10.18653/v1/2022.acl-long.445>. [17]
- Jurafsky, D. and J. Martin (2023), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Pearson Education Inc. [10]
- Koehn, P., F. Och and D. Marcu (2003), "Statistical phrase-based translation", *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, <https://doi.org/10.3115/1073445.1073462>. [21]
- Kustritz, A. (2015), "The Fan Fiction Studies Reader ed. by Karen Hellekson and Kristina Busse", *Cinema Journal*, Vol. 54/3, pp. 165-169, <https://doi.org/10.1353/cj.2015.0019>. [13]
- Manning, C., P. Raghavan and H. Schütze (2008), *Introduction to Information Retrieval*, Cambridge University Press, <https://doi.org/10.1017/cbo9780511809071>. [14]
- Mrini, K. et al. (2020), "Rethinking Self-Attention: Towards Interpretability in Neural Parsing", *Findings of the Association for Computational Linguistics: EMNLP 2020*, <https://doi.org/10.18653/v1/2020.findings-emnlp.65>. [11]
- OpenAI (2023), "GPT-4 Technical Report", *ArXiv [Cs.CL]*, <http://arxiv.org/abs/2303.08774>. [24]

- Palmer, M., D. Gildea and P. Kingsbury (2005), “The Proposition Bank: An Annotated Corpus of Semantic Roles”, *Computational Linguistics*, Vol. 31/1, pp. 71-106, <https://doi.org/10.1162/0891201053630264>. [7]
- Poesio, M. et al. (2016), *Anaphora Resolution*, Springer. [4]
- Rajpurkar, P. et al. (2016), “SQuAD: 100,000+ Questions for Machine Comprehension of Text”, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, <https://doi.org/10.18653/v1/d16-1264>. [18]
- Srivastava, A. et al. (2022), “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models”, *arXiv preprint arXiv:2206.04615*. [23]
- Sukthanker, R. et al. (2020), “Anaphora and coreference resolution: A review”, *Information Fusion*, Vol. 59, pp. 139-162, <https://doi.org/10.1016/j.inffus.2020.01.010>. [1]
- Wang, A. et al. (2018), “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, <https://doi.org/10.18653/v1/w18-5446>. [5]
- Wang, A. et al. (2018), “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. [6]
- Weischedel, R. et al. (2013), “Ontonotes release 5.0”, *Linguistic Data Consortium, Philadelphia, Pennsylvania*. [2]
- Xie, T. et al. (2022), “UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models”, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, <https://doi.org/10.18653/v1/2022.emnlp-main.39>. [20]

Annex 8.A. Natural Language Processing research areas

Annex Table 8.A.1. Natural Language Processing research areas with at least one benchmark task

	NLP research areas	Equivalent Human Language Competence
1	Anaphora resolution	Identify the anaphor for a specific antecedent e.g. the car (antecedent) is damaged but it (anaphor) still works.
2	Authorship verification	Identify the likelihood a text was written by a specific author or if a set of texts is likely written by the same author.
3	Automated essay scoring	Determine the academic quality of an essay.
4	Constituency parsing	Construct a constituency-based (or syntactic structure) parse tree for a sentence by applying phrase structure grammar rules.
5	Dialogue (open domain)	Carry out a conversation with another person about any topic.
6	Dialogue (task-oriented)	Assist someone in completing a task where the help provided is through conversation.
7	Emotion-cause pair extraction	Identify when a text refers to a person's emotional state and pair that with another location a text or conversation that describes the likely cause of that emotional state.
8	Event extraction / detection	Identify the events described within a text or in a social media post.
9	Explanation generation	Explain the reasoning that led to an answer.
10	Humour detection	Identify when a sentence or paragraph would be considered humorous to some/most people.
11	Image captioning	Describe what is shown in a photo.
12	Information Retrieval	Find (and rank) the content most relevant to a specific user information need.
13	Keyword extraction	Read a text and compose a set of keywords that are likely to be used as query terms when searching for that text in a large collection of documents, such as the web.
14	Lexical normalisation	Rewrite text written in non-standard form to standard form, e.g. Fomo? fear of missing out; lol? laugh out loud; jst? just.
15	Machine Translation	Interpret, translate text or spoken language.
16	Natural Language Inference / Sentence pair modelling (entailment, semantic, similarity, paraphrase detection)	Understand the relationship between the meaning of two sentences and how the meaning of one sentence can relate in some way to the meaning of another sentence.
17	Punctuation restoration	Add the most appropriate punctuation to a text for which punctuation is not present.
18	Question Answering	General knowledge or knowledge about a specific topic; Ability to find relevant information with the help of technology, e.g. search engine.
19	Reading comprehension	Reading comprehension including the ability to identify when the text provided does not include the information required to answer a question.
20	Relation extraction	Extract the relations between entities expressed within text or conversation, e.g. Michael Jackson died in Los Angeles, CA. where diedInCity is the relation.
21	Reasoning	Reasoning.
22	Semantic parsing/role labelling	Identify the semantic roles (e.g. agent, patient) for the arguments of the predicates (e.g. verb) of a sentence.
23	Sentence Compression	Rewrite a sentence as a shorter one by removing redundant information while preserving the meaning of the original sentence.
24	Sentiment Analysis (aspect-based)	Interpret the opinion being expressed within a text/conversation/social media post.
25	Sentiment Analysis (multimodal)	Interpret the opinion being expressed within multiple sources including language, e.g. facial expression and speech.
26	Speech recognition	Interpret spoken language.
27	Summarisation	Provide a written summary of a text or conversation that highlights the most important points made within a specified limit of words or sentences.
28	Text classification	Assign the text to the most appropriate category or domain (e.g. news, medical, scientific, etc.).
29	Text diacritisation	Restore the diacritics for text in languages that use diacritisation (Czech, French, Irish, etc.).
30	Text style transfer	Rewrite a text with the expressed content expressed in a specific style distinct from that of the original.
31	Topic modelling	Be able to identify the topic of a text/conversation/social media post.
32	Video captioning	Describe what is taking place in a video.
33	Visual QA	Answering a question about the content of a photo or image.

Notes

¹ Scores are an average of approximately 800 human ratings on a 100-point rating scale for a range of conversation quality criteria.

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^[11]). 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*NET¹ in the United States and ESCO² 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.

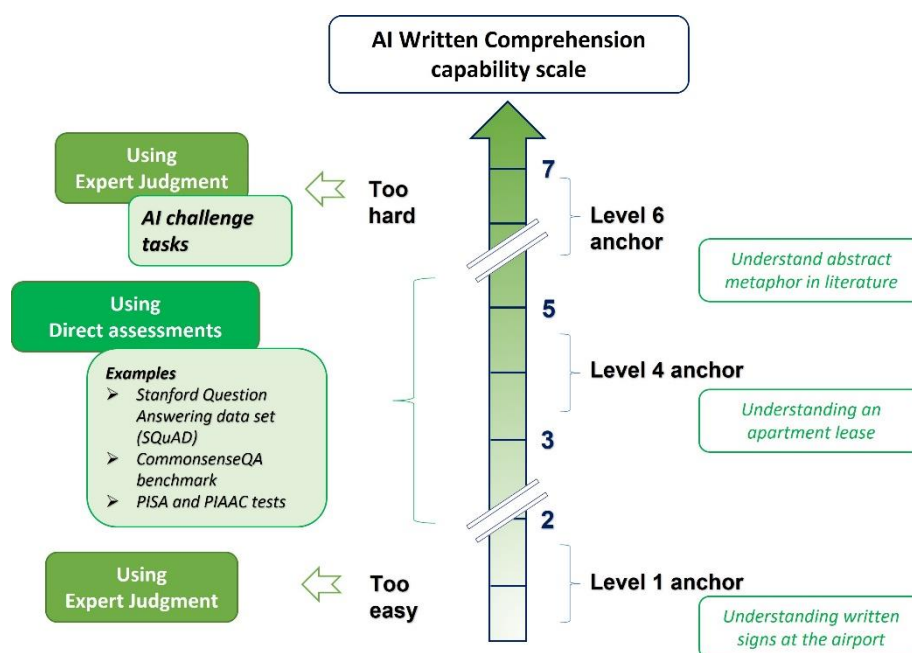
The selection of indicators to develop will reflect feedback from AI experts and policy makers about which taxonomy works best for translating AI capabilities to non-experts. It will also reflect their feedback on which scales from the occupational databases within that taxonomy are relevant to include 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).

Educational Research and Innovation

AI and the Future of Skills, Volume 2

METHODS FOR EVALUATING AI CAPABILITIES

As artificial intelligence (AI) expands its scope of applications across society, understanding its impact becomes increasingly critical. The OECD's AI and the Future of Skills (AIFS) project is developing a comprehensive framework for regularly measuring AI capabilities and comparing them to human skills. The resulting AI indicators should help policymakers anticipate AI's impacts on education and work.

This volume describes the second phase of the project: exploring three different approaches to assessing AI. First, the project explored the use of education tests for the assessment by asking computer experts to evaluate AI's performance on OECD's tests in reading, mathematics and science. Second, the project extended the rating of AI capabilities to tests used to certify workers for occupations. These tests present complex practical tasks and are potentially useful for understanding the application of AI in the workplace. Third, the project explored measures from direct AI evaluations. It commissioned experts to develop methods for selecting high-quality direct measures, categorising them according to AI capabilities and systematising them into single indicators. The report discusses the advantages and challenges in using these approaches and describes how they will be integrated into developing indicators of AI capabilities.



Federal Ministry
of Labour and Social Affairs



PRINT ISBN 978-92-64-88932-3
PDF ISBN 978-92-64-82429-4



9 789264 889323