19. Questions to guide assessment of artificial intelligence systems against human performance

Eva L. Baker, the University of California, Los Angeles

Harold F. O'Neil Jr., University of Southern California's Rossier School of Education

This chapter takes a step back from the project, reviewing practical issues around the assessment of artificial intelligence (AI) that must guide the next phase of research. It provides guidance for setting up a general analytical framework for assessing AI capabilities with regard to human skills. Questions are centred around three main issues. First, the chapter looks at how the various parameters of measurement depend on the objectives of the assessment. Second, it examines selection of tests for comparing AI and human skills. Third, it discusses selection and training of raters. The chapter concludes with a summary of issues considered in planning the study.

Introduction

Alongside the continuing explosion of artificial intelligence (AI) in research and applications is an accompanying need to understand its capabilities in a variety of settings. There is ample evidence that AI systems can solve particular problems far better and certainly faster than humans attempting comparable tasks. There is also a long list of AI accomplishments on sets of tasks humans could not directly attempt because of size and complexity of data. Nonetheless, there is a need to characterise and perhaps benchmark the capability of these systems.

This chapter discusses general precepts for consideration with respect to the planning study. It focuses on the issues surrounding the use or development of taxonomies and relevant measures to allow a new level of understanding of current and future capacity of AI systems. To that end, it poses questions and provides guidance related to the characterisation of AI systems, issues in organisation and task selection for assessment, and the rating process.

Why measure or characterise artificial intelligence systems?

This OECD project expects that measures of AI performance can be developed to compare to human performance based on a set of existing (or potentially newly constructed) measures. The approach tests the assumption that clear correspondences are possible for performance by AI systems and people. The project must clarify its needs before the best strategy and operational objectives can be proposed. Three large considerations are discussed below to stimulate reflection.

- Compare AI to human intelligence or to human workplace or academic outcomes?

Comparing AI to human intelligence versus comparing it to workplace or academic performances have implications for the tests and types of skills or tasks used for its assessment.

Based on the October 2020 workshop, the project is focused on the ability of AI to reproduce human performance in any conceivable human work or activity. To compare AI and human workplace performances, bounded domains of skills and content are much more useful than measurement approaches that depend on general constructs, e.g. math ability. However, to ensure that all human activities are covered, reflecting on the selected set of skills and content periodically is recommended. This helps map activities that were left out of the first selection (for example because they have emerged since the first assessment).

- Compare outcomes alone or combine outcomes and process?

AI systems use powerful computational approaches to solve problems or consolidate information. However, they can vary in terms of their strategy. Strategies include bottom-up exploratory approaches in machine learning, such as unsupervised learning, and a combination of rule-based and bottom-up data processing, such as expert systems.

The assessment may seek to understand at a conceptual level how AI systems accomplish their tasks rather than simply their success in solving problems or reaching conclusions. In this case, AI and human process performance could also be compared. Indeed, as all AI systems profit from their massive computer power to solve specific tasks, they may be subject to measurement error when a problem involves non-linear relationships, while humans may be able to adjust to such complexities. Hence, a focus on both process and outcomes will have implications for task selection for the study.

The OECD is primarily interested in outcomes, which is a suitable approach if AI produces the same outcomes with a different process. However, process is likely to be important for generalising the results of a test to a broader range of capabilities than are directly tested. Some generalisations may be valid for humans but not AI and vice versa. For example, common sense inferences can usually be assumed in

humans, while they must be tested explicitly in AI. That stems in part from process differences in developing expertise in AI and humans – AI expertise does not rest on a base of common sense the way it does for humans. Human-AI process differences should be kept in mind since they probably have implications for what needs to be tested on the AI side. Overall, some analyses or explanation of how the system arrived at its outcome must be provided for both acceptance and fairness issues.

- Document changes or improvements in AI competence over time or describe a single temporal point?

If the project intends to map the progress of AI systems across one or more sets of performance variables, then task selection should anticipate increases in AI performance over time. Thus, the tasks selected initially should represent areas in which AI currently does not do well. This would avoid a ceiling effect for later measurement occasions. If the intention is for one time only, then a range of "difficulty" or complexity of tasks could be included to detail how the systems operate today.

Issues in organisation and task selection

Selection of organisational systems and the corresponding tasks are major concerns for the study. The questions below amplify or posit additional prerequisite concerns about the best ways to determine taxonomies or organisational structures. Should the project select existing ones or develop new ones? The questions also examine assessments to be considered by the project.

Context of assessment tasks

- How important are the contexts of use of the assessment tasks? Is the project considering tasks drawn primarily from an academic disciplinary base or tasks that include context, such as general and specific knowledge?

If the interest is in disciplinary tasks, such as tasks of mathematical concepts or interpretations of the meaning of texts, then the project might assemble data from an age- or experience-based range of subjects. It would then ask the project rating team to make their estimates of AI systems using categories that capture span of respondent differences. For example, raters might be asked to estimate how well upper elementary students, junior high school algebra pupils or college math students could solve a problem. This distribution of performers would provide a database to guide the raters in making their judgements. However, disciplinary assessments often include implicit contextual information. This could have unintended effects on performance, especially if background characteristics of respondents are unknown or not included in the study.

Instead of disciplinary-focused tasks or items, tasks could be selected from areas in workforce skills. While there is contextual and background knowledge here, much of it is job related, which the systems could learn. This strategy is recommended if the project wants to influence how AI is perceived, used or evaluated in workplace context. It involves the selection of tasks (both comparable and unique) from two or more sectors, e.g. finance, hospitality. This method would indicate the robustness of a system across sectors and types of embedded domain knowledge. It could also give useful information for similar tasks, as well as about the functionality of the system on tasks wholly unique to a particular sector (think of a Venn diagram showing shared and unique knowledge and skills).

Prescriptive vs descriptive functions

- Is prescription or description intended as the principal function of the taxonomy or other organising system with respect to the experimental tasks? Is the taxonomy to guide selection of tasks, i.e. to

function prescriptively? Or, will the organising system reflect, in a verbal or symbolic way, the nature of the tasks to describe and represent them as they exist?

If tasks were selected, perhaps from domains assessed by the German vocational/technical system – in two or more sectors – what similarities and differences in assessment design structure would be found? How might such bottom-up analyses among sectors relate to the O*NET structure? Depending on the sphere of influence selected as the major target of the project, a hybrid approach could be adopted. This approach would use formulations that guided the creation of tasks, i.e. German sector tasks or O*NET descriptors. It would then modify organisational structures and descriptions depending upon whether the purpose was an existence-proof project or the beginning of a sequence of connected studies about AI progress.

The OECD has been thinking of the taxonomies functioning prescriptively to guide task selection for two reasons. First, taxonomies can serve as a reminder of capabilities or workplace activities that need to be assessed. Second, they can point to capabilities or workplace activities that are closely related and should be assessed together. A descriptive approach would make sense if specific work contexts are of interest, each with well-defined authentic tasks that would provide the natural focus of assessment. In that approach, the OECD could see the need to make sense of the various tasks that arise in those contexts. However, in principle, the full range of current and possible work contexts must be addressed. Therefore, a prescriptive framework will more likely be needed. This framework would consist of taxonomies of capabilities and workplace activities to use in specifying what needs to be assessed.

Characterising artificial intelligence systems

- How can AI systems be characterised?

Following the selection of tasks and tests, experts will be selected to rate how computer systems would fare with the problems or tasks. This can be done in multiple ways.

First, through procedures that permit the direct processing of tests or tasks by existing AI systems themselves.

Second, by assessing the capabilities of specific AI systems. This approach may yield more reliable judgements but involves some decisions:

  - Which system attributes will be included? How will they be represented or described? Will experts be expected to understand a subset of common AI systems?
  - Should AI systems that have an explanatory function be selected or rather those that do not (e.g. a white box or black box)?

Third, by assessing the capabilities of current technology in general. In this case, experts rate the feasibility of well-defined tasks based on their knowledge of state-of-the-art AI approaches.

There are many ways to combine and describe data for this project. For example, there are several options regarding how judgements are made, how ratings are tabulated and summarised. Rating can take place in various forms such as through interface development options.

Rating process

This section suggests refinements to the human rating process. These approaches are derived from the extensive experiences available from ratings of open-ended achievement tests, as well as from new ideas stimulated by the workshop. All decisions regarding rating have cost implications, to be discussed in a subsequent section. Cost implications interact with decisions to have one status study or a sequence of studies over time assessing AI system progress.

Raters

Raters need to understand the nuance, as well as the operations, of their study tasks. They must have both adequate training and time to accomplish their judgements. Finally, the processes they used must be documented.

Number and selection of raters

- Who will be the actual raters of the assessment materials? Who will determine if the selected tasks are representative of the desired domain or sub-domain? How will the tasks be identified and selected?

Published tasks may or may not be thoroughly vetted. Unless there is robust evidence on this process, a group of experts should examine and judge their quality and suitability for the study.

A second group of raters, the AI experts, will then judge how well AI systems could complete all or part of any tasks.

For both rater groups, explicit criteria should be set for selection, such as expertise, availability (based on notions of ideal timing for the project, estimated time for training and the study ratings, post-rating interviews, willingness to participate in subsequent ratings).

- How many raters are needed?

This decision depends first upon the estimated numbers of tasks or items to be rated. If the tasks are selected from a common disciplinary area, such as mathematics, then raters should be at least familiar with the area and have teaching experience with the content of the tasks. One would also want raters to demonstrate competence on the study tasks. The project may wish to complete its work in a single language, e.g. English.

- Should raters be selected from homogenous work environments or represent a broader swath of backgrounds and experience?

On the one hand, irrelevant task variance should be limited. On the other, a set of raters with a limited or perhaps peculiar orientation based on the source of their experience should be avoided.

- Could an English-only study be replicated in other OECD countries?

Bilingual raters could ensure the same level of stringency is used in any between-country study. If multilingual raters are used in the main study, a minimum of three from each country could be used for their topic.

Expectations of performance would vary by country largely in any domain as a function of nationality. For example, American and Japanese raters would have distinct expectations for their own students' expertise in given subject matters. Psychologists and measurement experts would rate quality of items along a set of dimensions. However, like the subject matter experts, they will be influenced by their experiences.

Instead of pure measurement experts who likely have preferred methods if not ideology, it may be best to find experts with a "conjoint" interest in learning and assessment. In this instance, conjoint refers to individuals who understand the learning requirements for success at particular tasks and who can judge item or task quality as well. It does not mean formative assessment devotees who advocate specific classroom assessments for teachers. Experts from the learning sciences would be ideal.

Work of raters

- How will raters work, where and for how long?

Ratings could take place at a common site or distributed sites, e.g. schools or homes. Depending upon timing, training and rating can take place over Zoom or other platforms that permit recording. In rating A-Level exams in England, raters experienced interruptions. Compliance among individuals varied. Some completed all tasks as rapidly as possible during a single sitting despite their given instructions to complete a certain number (randomly ordered) each day.

Creation of scoring rubrics

- What kind of supports do content and AI raters need for the rating process?

The process should avoid binary responses (e.g. where AI either can or cannot successfully address the item or task). An agreed-upon protocol could constrain the AI raters to a specific range of systems. Either it would provide system functional descriptions or architecture, the kinds of output or uses to which the system has been put, or use a standard set of AI systems. This last choice would require a qualifying "test" to determine that prospective AI raters know and understand the systems to be considered.

Both rater groups will need rubrics to guide judgement of analytical components, as well as to help them make an overall rating of the likely success of the system (or system types, if more than one is considered). There have been many studies of rubric construction, reliability, clarity and other characteristics. The rationale behind experts' judgements should be documented.

Ideally, rubrics could begin with an overall judgement on a preferred scale, e.g. 1-4, 1-6. This would be followed by a sub-scale scoring of attributes or elements that may have been considered in that decision. There has been work that begins with sub-scores and adds them up for a final judgement, but that practice turned out to be less useful.

In addition to an overall judgement, sub-elements on the rubric might describe a comparable task known to have been successfully encountered by the system, the logic or linearity of the task, relevance of specific prior knowledge, the clarity of the language used in the task, and so on.

For the rubric used to judge test tasks or items, subcomponents could relate to the importance of the tasks, the critical content or contextual elements it includes, its dependence on prior knowledge, clarity of task for the examinee, and – if open-ended – the quality of the scoring rules intended to be applied.

- How much detail should rubrics have? How should they be developed?

Most experience advises against long and detailed scoring rubrics because of fatigue, lack of inter-rater agreement and the general inter-correlation clusters of elements.

It is good practice to create and try out rubrics before training raters to use them. In that process, a separate set of experts is given a set of experiences that model the range and timing of a single scoring session. Raters may independently use the rubric and note difficulties. Other approaches involve some "think-aloud" protocols where the raters speak aloud the process they are using to arrive at scores for both the overall and sub-categories.

Judgements about rubric quality tend to cluster around its clarity, applicability to the range of tasks, brevity and understanding of the meanings of rating points, (how does a "3" differ from a "4?").

The development of rubrics should also involve verifying that experts obtain similar ratings on the same tasks or items. Rubric development experiences should ideally be sequenced. This allows the rubric to be tried on two or so items, notes and scores compared, and discrepancies found and discussed. The rubric can then be revised and used for the next small set.

The development process should also examine whether scores are more sensible when the raters score as they go, or when they first only make notes and score all elements at the end. This latter approach can ensure that the same scale values are used across rating sub-topics. Once the process has stabilised the

rubric, a trial rating occurs to estimate the time it takes to score. Limiting scoring time per task is good practice.

Rater training

- How should raters be trained?

Rating training could consist of a face-to-face work (preferred) or a combination of face-to-face and webinar with feedback. Rater training involves both training and qualification. Training involves using the same rubric applied to a *different* set of examples than those to be encountered in the actual study.

Three sets of technical quality criteria are relevant. First, raters need to agree with a criterion set of ratings, using approaches that cleave to the rubric-incorporating experts' views. Second, there should be intra-rater agreement, meaning that a single rater assigns much the same value to identical or similar tasks over time. Third, there should be inter-rater agreement on the same task, usually expressed as the obtained percentage of exact agreement between pairs of raters. In practice, the first two criteria are more important.

The order of rating needs to be organised to represent a number of elements. A general random order with repeated examples (probably three) should establish consistency within raters. There should also be a way to enter scores and have computations of agreement take place on the spot.

If agreement is found to be low, following the ratings, the "leader" can discuss the ratings in terms of the criterion cases, but there should be no opportunity for raters to discuss and modify their scores. Such discussion usually results in a socially defined rating that may be unique to subsets of raters and undermines the validity of the process.

Finally, more individuals than needed will need to be trained because some raters will likely not qualify. Ideally, another project-related task may be found for them, such as compiling and summarising information.

Answering the questions

The OECD project posed six questions at the outset, which are answered below.

- How should the project decide on what types of skills to assess?

Domains with sets of coherent knowledge, skills and attributes are best for tasks related to performance using skills and content (rather than broad constructs). The German efforts on tasks and standards within the world of work sectors are relevant and of high quality, both to the project and to AI system development. Raters could also be given additional tasks that will stretch even the best of current system capabilities (transfer). This is because elements of transfer could be described related to completeness of queries, types of logic, extrapolations and so on.

- Are there existing skills taxonomies that describe the right set of skills to assess?

If a workplace setting is chosen, then the skill taxonomies underlying the German or other such system (like O*NET) should be used.

- How should the project decide on the types of tests to be used?

One or more bounded domains in a clear use context could be used as the source of assessments. This approach will have a greater likelihood of success because it will have limited abstraction and communicate best with desired audiences. Operationally, the use of a compilation of tasks or items rather than selecting an existing test is advised. There will be trade-off between relativity and validity in either case.

At least two sectors or large content domains are advised to demonstrate generalisability of the approach. In addition, within a domain, a range of task and item types should be employed, probably no less than three for each type. Common descriptors of items will include item formats, ranging from brief to extended, that provide or expect recollection of essential prior knowledge, which include selected or open-ended responses. Open-ended responses should include multi-stepped solutions to problems that are ill formed, partially formed or where the problem statement is given. Tasks that require various iterations among task formulation, resource search and acquisition, integration and solution fit are desirable.

Assuming systems will have natural language processing capability, there may be excellent tasks that require paraphrasing. This will create a text answer that demonstrates understanding of the levels of meaning of particular questions that are independent of any specific content domain.

Items used should vary in complexity and completeness. Ideally, data should be available regarding item difficulty or facility with answers for different groups. Thus, items and tasks to be selected should include a range of cognitive requirements, task formats, and use of language and other underlying skills.

- Are there existing types of tests that are good candidates for assessment?

The work provided at the October 2020 workshop for the *Artificial Intelligence and the Future of Skills* project would be a good start. Self-report and game/simulation measures are available from previous research.

- Are there existing types of tests that should be excluded from the assessment?

Well-known standardised tests for admissions to higher education or for evaluation and accountability in pre-collegiate environments should be avoided. Some communities see these tests as biased.

- Are there new types of tests that need to be developed for the assessment?

Yes, if both outcomes and processes can be identified for this environment. Rather than a completely new development, it seems likely existing tests could be modified.

Further reading

AERA, APA and NCME (2014), *Standards for Educational and Psychological Testing,* American Educational Research Association, American Educational Research Association, American Psychological Association and National Council on Measurement in Education, Washington, DC.

Baker, E.L. et al. (2022), “Assessment principles for games and innovative technologies”, in O’Neil et al. (eds.), *Theoretical Issues of Using Simulations and Games in Educational Assessment,* Routledge/Taylor & Francis.

Baker, E.L. et al. (2019), *Validity Studies and Noncognitive Assessments* (Deliverable Item No. 013 to funder), National Center on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Bergner, Y. and A.A. von Davier (2019), “Process data in NAEP: Past, present, and future”, *Journal of Educational and Behavioral Statistics*, Vol. 44/1, pp. 706-732.

Buhrmester, M.D., S. Talaifar and S.D. Gosling (2018), “An evaluation of Amazon’s Mechanical Turk, its rapid rise, and its effective use”, *Perspectives on Psychological Science*, Vol. 13/2, pp. 149-154.

Choi, K. et al. (2021), *Molly of Denali Analytics Validation Study Report—final* (Deliverable to PBS KIDS)., Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Chung, G.K.W.K. (2015), “Guidelines for the design, implementation, and analysis of game telemetry”, in C.S. Loh, Y. Sheng and D. Ifenthaler (eds.), *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, Springer, New York.

Clariana, R. and E. Taricani (2010), “The consequences of increasing the number of terms used to score open-ended concept maps”, *International Journal of Instructional Media*, Vol. 37/2, pp. 163-173.

Dunbar, S. B. and C.J. Welch (2022), “It’s game day in large-scale assessment: Practical considerations for expanded development and use of simulations and game-based assessment in large-scale K-12 testing programs”, in O’Neil, H.F. et al. (eds.) *Theoretical Issues of Using Simulations and Games in Educational Assessment,* Routledge/Taylor & Francis.

Kirkpatrick, J.D. and W.K. Kirkpatrick (2016), *Kirkpatrick’s Four Levels of Training Evaluation*, Association for Talent Development, Alexandria, VA.

Klein, D. et al. (2002), *Examining the Validity of Knowledge Mapping as a Measure of Elementary Students’ scientific understanding* (CSE Tech. Rep. No. 557), National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Mislevy, R.J. et al. (2015), “Psychometrics and game-based assessment”, in C.S. Loh, Y. Sheng and D. Ifenthaler (eds.), *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement,* Springer, New York.

O’Neil, H.F. et al. (2021), *Using Cognitive and Affective Metrics in Educational Simulations and Games: Applications in School and Workplace Contexts,* Routledge/Taylor & Francis.

O’Neil, H.F. et al. (1994), “Human benchmarking for the evaluation of expert systems”, in O’Neil, H.F. Jr. and E.L. Baker (eds.), *Technology Assessment in Software Applications*, Hillsdale, MI.

Quellmalz, E. et al. (2012), “Science assessments for all: Integrating science simulations into balanced state science assessment systems”, *Journal of Research in Science Teaching*, Vol. 49/3, pp. 363-393.

Schenke, K. et al. (2021), “Measuring and increasing interest in a game”, in O’Neil, H.F. et al. (eds.) *Using Cognitive and Affective Metrics in Educational Simulations and Games: Applications in School and Workplace Contexts,* Routledge/Taylor & Francis.