18. Tasks and tests for assessing artificial intelligence and robotics in comparison with humans

Art Graesser, Psychology and the Institute for Intelligent Systems, University of Memphis

This chapter reflects on major issues recounted in the previous chapters, raising three key questions relevant to the *Artificial Intelligence and the Future of Skills* project. What is the value in identifying ideal models when comparing humans with artificial intelligence (AI) and robotic systems? How might systematic mapping occur between skill taxonomies, tasks, tests and functional AI components? How can major differences be handled in targeted skills, different occupations and changes in the world? Some suggestions are offered on next steps in addressing these questions.

Introduction

The OECD *Artificial Intelligence and the Future of Skills* (AIFS) project attempts to understand the educational implications of artificial intelligence (AI) and robotics. The goal is to create an ongoing programme to assess the capabilities of AI and robotics, and to compare them with human capabilities. The 5-6 October 2020 meeting featured presentations from 13 experts who spanned diverse fields, including educational assessment, AI, robots, cognitive science and workforce training. Additional experts from various fields (and countries) also contributed. This chapter raises some questions and suggestions, and offers other reflections on major issues covered at that meeting and recounted in the preceding chapters.

Identifying knowledge, skills and abilities

One central issue is to identify the set of knowledge, skills and abilities (KSAs) to assess. Psychology has proposed comprehensive taxonomies with psychometric *tests.* These include the three-level Carroll-Horn-Cattell model presented by Kyllonen (see Chapter 3). This has a long history of validation in humans and quantitatively tuned factor analyses.

There are abilities and skills identified in industrial-organisational psychology and business that involve *tasks* specific to particular occupations. This allows adults to be trained and certified to practice in the occupation. For example, Dorsey and Oppler (see Chapter 10) described the O*NET (Occupational Information Network) in the US Department of Labor. It identifies KSAs for occupation categories (e.g. manufacturing, health care). Rüschoff (see Chapter 9) presented the vocational education and training framework in Germany. It has an intense two-day assessment that has practical, written and oral components, including answering questions to justify actions.

Greiff and Dörendahl (see Chapter 7) pushed the envelope beyond basic cognitive skills and domain-specific skills into the realm of *transversal* skills that have increasing importance in the 21st century. These comprise problem solving, collaboration, creative thinking and global competency. Wooley (see Chapter 6) echoed the importance of social intelligence and collaboration. Conversely, De Fruyt (see Chapter 5) emphasised social skills and emotion regulation skills.

The AI/robotics contingency did not offer taxonomies of KSAs, as pointed out by Hernández-Orallo (see Chapter 11). Instead, it focuses on *functional components* of intelligent mechanisms, such as knowledge representation, reasoning, planning, learning, perception, navigation and natural language processing. They evaluate how well the various computational models in AI/robotics compare with humans on tasks that focus on these functional components.

Forbus and Davis (see Chapter 2 and Chapter 12, respectively) pointed out which components are easier for computers to achieve (such as remembering and accessing facts) and which are easier for humans (such as common sense reasoning). Avrin (see Chapter 15) discusses systematic evaluations of over 900 AI systems on recognition capabilities, learning, understanding, generation and mission navigation.

Nearly all of these evaluations of systems in AI/robotics have been on practical tasks. Such tasks, such as autonomous cars and text summarisation, have objective criteria of success. Moreover, the tasks typically focus on those performed by adults in the workforce. However, Cheke (see Chapter 17) covered low-level skills of animals whereas Chokron (see Chapter 4) focused on cognitive and social skills of children. These presentations address the second central issue of the expert meeting, namely identifying differences in what can be accomplished by humans versus AI/robotic systems.

With this context in mind, the chapter raises three questions, with associated reflections and suggestions. These aim to shed light on the primary goal and two central issues of the AIFS project.

What is the value in identifying ideal models when comparing humans and artificial intelligence/robotic systems?

Towards a formal model of performance to assess and compare humans and AI robots

One could imagine an ideal specification (i.e. formal model) of performance on tasks, tests and functional components. Such a model could serve as a standard to assess and compare humans versus AI/robots. That would go a long way in providing a fair comparison on the capabilities of the two systems.

A perfect ideal model is perhaps illusory, but there can be approximations. For example, accomplished human experts can specify ideal responses to tasks and tests. These could either solve a problem or meet a level of mastery in achieving particular tasks. Such a specification has both a content analysis and a threshold analysis.

Content analysis

The content specification would declare the particular behaviours and products that correspond to a successful accomplishment of a task. This approach is adopted by designers of intelligent tutoring systems (Koedinger, Corbett and Perfetti, 2012[1]; Graesser, Hu and Sottilare, 2018[2]). These systems identify *knowledge components* required to master a subject matter or skill (e.g. algebra, physics). They also prepare a *Q-matrix* that specifies the knowledge components associated with each particular problem, task, or item along with behavioural manifestations of each knowledge component mastery. A complete and accurate solution would be needed, but it might also consider intermediate levels of achievement.

Threshold analysis

While the content analysis is applied to each individual item on a test or step in a task, the threshold analysis is applied to an aggregate score from the entire test/task. The threshold analysis identifies points on a continuum of scores that predict practical external criterial outcomes (which is infrequently conducted). This contrasts with exclusively psychometric indices or breakpoints in the distribution of scores (which is routinely conducted). Analyses can assess how well a population of humans or AI/robot systems meet the various thresholds of scores.

How well would a human vs. an artificial intelligence/robotics system perform?

Each adult has decades of experiences to fortify them in a task. Information about this past is either non-existent or minimally specified through demographic data or assorted tests. AI/robotics systems are unlikely to have such data available. However, there are AI systems that learn with experience. The *Never-Ending Language Learner* (NELL), for example, runs 24 hours per day learning to read the web and grow a knowledge base of beliefs (Mitchell et al., 2018[3]).

There are several possible approaches to understanding how AI/robotics system can be put on an even playing field with a human. The first approach assumes AI systems and humans have a different array of assets and resources, and thus are rarely on an even playing field. However, they can still be compared on tasks, which the AIFS project is planning. A second approach puts the system through a practice set of benchmark tasks for a month. It then grades performance on a test set, as in the case of the NIST methodology. A third approach is to conduct an AI/simulation over a long stretch of time or epochs of experiences. The performance produced in such tasks is then observed, as in NELL (Mitchell et al., 2018[3]).

A computational or information-theoretic analysis could specify a problem space, combinatorial landscape or another type of formal, quantitative model that identifies hypothetical alternatives and bone fide

solutions. For example, the iconic "travelling salesman problem" attempted to find a route between 40 cities that minimised the distance in travel time. The problem proved to be so hard that it would have required over 1 000 years on the fastest computer that existed 20-30 years ago. Who knows how the travelling salesman problem fares 30 years later? However, computational analyses like these can be posed for a fair comparison between humans and AI/robotics.

Ideal models with both computational and human constraints

There are ideal models that incorporate both computational and human constraints. For example, cognitive scientists often perform tasks analyses on particular problems or problem sets that decompose the solution plans and concrete steps in executing solutions (Anderson, 2009[4]; Laird, 2012[5]). Researchers can compute the probability and time of (sub)task completion, as well as the assorted solution strategies.

A good example of this approach is the models of lower-level perceptual-motor tasks. These include the Goals, Operators, Methods and Selection Rules (GOMS) model (Card, Moran and Newell, 1983[6]) and CogTool (John, 2013[7]). A researcher first specifies a set of tasks and the model generates expected task completion times and other aspects of performance.

GOMS and CogTool are based on an ideal rational model (much like Anderson's ACT-R and Laird's SOAR) and psychological components (such as perception-cognition-action cycles, production rules) and psychological laws. Fitt's law, for example, computes the time to move a part of a body to a target. Hick's law specifies that the time taken for a decision is a logarithmic function of the number of alternatives. The power law of practice specifies an exponentially decreasing function of task completion time as a function of number of attempts to complete a task.

GOMS and CogTool are remarkably accurate in predicting performance in some tasks (Graesser et al., 2018[8]) but not others that require higher-order reasoning. Consequently, GOMS is best used to complement rather than replace expert judgements of task difficulty. Nevertheless, for lower-level tasks involving well-practised procedures, researchers would have a foundation for comparing humans and robots. Similar approaches could be proposed for problem solving, reasoning, collaboration and other transversal skills (Sinatra et al., 2021[9]).

How might systematic mapping occur between skill taxonomies, tasks, tests and functional artificial intelligence components?

Participants at the expert meeting presented several skill taxonomies, tasks, tests and functional AI components. A few of these are presented below.

- Carroll's (1993) 3-Stratum model presented by Kyllonen (Chapter 3)

As discussed in Chapter 3, Stratum 3 is a general intelligence factor, whereas Stratum 2 has eight constructs manifested in factor analyses: fluid intelligence, crystallised intelligence, general memory/learning, broad visual perception, broad auditory perception, broad retrieval ability, broad cognitive speediness and processing speed. Stratum 2 depends on the measures collected in Stratum 1, which consists of dozens of precisely operationalised tests from the field of psychological assessment (Carroll, 1993[10]).

- Greiff and Dörendahl's (Chapter 7) taxonomy (which includes transversal skills)

There is a distinction between transversal skills (problem solving, collaboration, creativity, digital competence, global competence), core domain skills (mathematical, reading and science literacies) and basic cognitive skills (general mental ability, fluid reasoning, comprehension knowledge, working memory, and others discussed in Chapter 3).

- Cheke, Halina and Crosby's taxonomy (which emphasises lower-level cognitive skills in both animals and humans)

Object and space skills include spatial memory and navigation, object representations and causal reasoning. Social and communicative skills include social learning and communication and social cognition, with several behavioural tests or tasks to operationalise these major categories.

- Hernández-Orallo's functional components in AI/robotics

The functional components in the list were knowledge representation, reasoning, planning, learning, perception, navigation and natural language processing. However, there may be others as AI/robotics evolves.

There were other taxonomies and distinctions discussed in the expert meeting. These included emotion regulation, empathy, trust and other dimensions of human experience (smell was an intriguing example). Perhaps the detection of misinformation would be particularly relevant in the age of social media (Rapp and Braasch, 2014[11]). In theory, there are subcomponents of misinformation detection, such as identifying the expertise of the source of information, comparisons to information in other documents, status of the media outlet, and sophistication of the language or information delivery.

Many skill categories and distinctions are potential candidates. Consequently, there are challenges in identifying which skills to include. One could adapt at least four approaches to meeting the challenges.

- The *comprehensive approach* would include any skill included by two or more stakeholders in the project, noting the lack of impact of unusual singletons.
- The *consensus approach* would include those skills that a sufficient number of stakeholders would endorse.
- The *intersection approach* would include those skills that can be measured in both humans and AI/robots in action.
- The *theoretical approach* would adopt one singular model for all to adhere to.

Approaches to selecting functional components

The *intersection approach* is compelling because the measures are available. This means it would be pragmatically strategic to implement them. However, this approach would need to consider comprehensiveness and the theoretical landscape. Perhaps the most pragmatic solution is to identify a small number of skills/tasks that represent different areas of the theoretical landscape.

A *consensus approach* would require a mapping between the different categories in different taxonomies, as exemplified above. Such a mapping could facilitate understanding of the landscape of skills, but major difficulties will emerge. For example, reasoning in one taxonomy will mean something different in another taxonomy. To address these concerns, stakeholders would need to negotiate a common ground of conceptual meanings of constructs, operational definitions of measures and other considerations. These would need to be familiar to those in the world of assessment, as well as science more generally.

Differences in semantics will occur between different research communities and stakeholders. For example, mundane plausible reasoning in human taxonomies is different from the formal reasoning in propositional calculus, AI's theorem proving and even the inference rules in the CYC computer system that represents world knowledge (Lenat et al., 2010[12]). Humans do well on *modus ponens* (If X, then Y; X; therefore Y). They consistently fail on *modus tollens* (If X, then Y; not Y; therefore, not X). Further, they often embrace the abductive reasoning that has an illegitimate formal foundation (Rips, 1994[13]). Similarly, statistical reasoning is different in formal systems vs. humans. Humans are prone to have, for example, base-rate and hindsight biases (Kahneman, Slovic and Tversky, 1982[14]). Such facts, of course, have relevance to what humans versus AI/robot systems can accomplish. That is apparent and also interesting.

In sum, reasoning is different in the various taxonomies. These differences reflect the goals of the projects, differences in fields and history. Negotiations among relevant stakeholders will be necessary to converge on a common ground. The achievement of a deep mapping of categories between taxonomies is complex and perhaps unlikely at the fine-grain levels but routinely successful at a course-grain level.

There are caveats, of course. The achievement of a loose mapping of categories is easier but possibly misleading because of non-trivial differences that end up getting missed. A small number of broad categories risks glossing over major differences in specific tasks/tests selected to represent the broad categories. Mapping between taxonomies is thus beset with serious challenges, but the history of assessment offers encouragement that the goals can be pragmatically achieved.

Mapping taxonomies and tasks, tests and functional artificial intelligence components: the Q-matrix

It is essential to generate mappings between particular taxonomies and the specific tasks, tests and functional AI components. In some circles, these are called a *Q-matrix*. Each item (e.g. question to answer, alternative to select, action to perform) in an evaluation scenario is assigned a code of attributes being assessed by an item (i.e. knowledge component, knowledge, skill, strategy, ability).

There can be primary, secondary and tertiary codes in these expert annotations. Stakeholders from different professional communities can annotate the items in a candidate scenario on the taxonomy categories of importance. The analysts in each stakeholder community could adopt whatever standards and criteria they wish to adopt, as long as other stakeholders can understand them.

What can be accomplished with the Q-matrices at hand from various stakeholders on candidate items in scenarios? The different stakeholders can evaluate and give feedback on whether the scenarios and items have a sufficient representation of the important taxonomic categories in their community.

Just as countries give such feedback in OECD international assessments (e.g. the Programme for International Student Assessment), stakeholders from relevant communities can give their feedback. Approval of scenarios and tasks depends on constraint satisfaction and negotiation. Relevant stakeholders need input on most phases of the assessment – from selecting relevant scenarios and tasks to developing illuminating items with the associated constructs they manifest. Items in this context may be actions in addition to verbal contributions and decisions in conventional assessments.

How can major differences be handled in targeted skills, different occupations and changes in the world?

The tasks, tests and functional components under focus are different. Occupations have different expectations. Subject matters are different among the occupations. The world also changes in trajectories that differ among countries, languages and cultures. How can these differences be accommodated in AI systems?

Consider separate implementations for each occupation, skill and time slice

As one simple answer, AI will need separate implementations for each occupation, skills and time slices being considered. This can be accomplished surprisingly quickly if certain conditions are met:

- a sufficient corpus of data for training and testing with machine learning
- a sufficient crew of knowledge engineers for annotation of data (needed for supervised machine learning) and development of scripts, rules or other modules with authoring tools.

This would require funding. However, it is a matter of availability of resources, engineering and investments as opposed to a devasting bottleneck. Whether general AI principles and mechanisms can be gleaned from such activities is an open question.

Compare timepoints

It is, of course, important to address bias in many of these questions, as well as changes that occur over time. As articulated by Greiff and Dörendahl in Chapter 7, there is a shift in the need for transversal skills. Therefore, problem solving, collaboration, reasoning and creativity in the world will have a higher impact on predicting the successful workforce profile than will memory and routine perceptual-motor skills. The workforce data clearly reveal this shift (Autor, Levy and Murnane, 2003[15]; Elliott, 2017[16]).

It could be argued that AI/robotics systems have not made significant headway in self-regulated activities and many of the transversal skills. This puts them at a disadvantage in these 21st century KSAs in contrast to their clear superiority in retrieving facts. Nevertheless, these points are non-problematic. For now, the goal is to identify what skills can be accomplished by humans vs. AI/robotics systems.

Some comparisons between timepoints might help project the workforce of the future and assess generalisation of claims. In one approach, a collection of scenarios and tasks is representative of the past, vs. the present vs. the future in the ultimate assessment. That is, a subset of the scenarios would represent the world of ten years ago; another subset the present; and another subset the uncertain science fiction of the future.

The three time points would crudely track trends over time on the measures collected from individuals at different age partitions and occupations. To increase the precision of temporal trends, time can be divided into finer slices. It would move from the past through present so that linear and non-linear trends can be detected and projected according to different quantitative models. However, projections would be considered with caution. Revolutionary disruptive historical changes periodically occur, such as war, pandemics like COVID-19 and the escalation of technology.

The selection of assessment scenarios and items will need to accept how existing tests, tasks and functional AI components have a distinctive history that may resist compromise. Perhaps the assessment materials that end up being selected/created will be a blend of the different traditions. In this way, they may have a chance to pacify multiple stakeholders. Perhaps the selected assessment scenarios will be fortified by Q-matrices that have tentacles to most or all of the stakeholders. Whatever scenarios end up selected, systematic comparisons will be needed between humans and AI/robotic systems.

Recommendations

- **Use ideal models as a neutral standard**

Some ideal models could serve as a neutral standard in comparisons of AI/robotics systems and humans. Ideal models are likely to stimulate exciting research on the data that end up being collected. However, they could potentially influence the selection of scenarios/tasks/tests.

- **Adopt an intersection approach to select tasks/tests for the comparisons**

The intersection approach uses tasks/skills/components that have been investigated both in psychology and AI, and covers different regions in the theoretical landscape. These decisions will require negotiations among stakeholders, as has been routinely accomplished for decades in the world of assessment.

- **Select scenarios, tasks and skills that present past, present and future**

The tasks performed in the work and daily lives of adults are known to vary over decades. Therefore, it would be prudent to select scenarios, tasks and skills that are representative of the past, present and

future. This would permit detection of trends over time for participants in different age groups, occupations and demographic characteristics. However, projections must be tempered with caution to the extent there are disruptive historical events such as COVID-19.

References

Anderson, J. (2009), *How Can the Human Mind Occur in the Physical Universe?*, Oxford University Press. [4]

Autor, D., F. Levy and R. Murnane (2003), "The skill content of recent technological change: An empirical exploration", *The Quarterly Journal of Economics*, Vol. 118/4, pp. 1279-1333, https://doi.org/10.1162/003355303322552801. [15]

Card, S., T. Moran and A. Newell (1983), *The Psychology of Human-computer Interaction*, Lawrence Erlbaum Associates, Hillsdale, NJ. [6]

Carroll, J. (1993), *Human Cognitive Abilities: A Survey of Factor-analytic Studies*, Cambridge University Press, New York. [10]

Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264284395-en. [16]

Graesser, A., X. Hu and R. Sottilare (2018), "Intelligent tutoring systems", in Fischer, F. et al. (eds.), *International Handbook of the Learning Sciences*, Routledge, New York. [2]

Graesser, A. et al. (2018), "Via: Using GOMS to improve authorware for a virtual internship environment", in Roscoe, R., S. Craig and I. Douglas (eds.), *End-user Considerations in Educational Technology Design*, IGI Global. [8]

John, B. (2013), *Cogtool (Version 1.2.2) (Software)*, https://github.com/cogtool/cogtool/releases/tag/1.2.2. [7]

Kahneman, D., P. Slovic and A. Tversky (1982), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, New York. [14]

Koedinger, K., A. Corbett and C. Perfetti (2012), "The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning", *Cognitive Science*, Vol. 36/757-798, http://dx.doi.org/10.1111/j.1551-6709.2012.01245.x. [1]

Laird, J. (2012), *The SOAR Cognitive Architecture*, MIT Press, Cambridge, MA. [5]

Lenat, D. et al. (2010), "Harnessing Cyc to answer clinical researchers' ad hoc queries", *AI Magazine*, Vol. 31/3, pp. 13-32, http://dx.doi.org/10.1609/aimag.v31i3.2299. [12]

Mitchell, T. et al. (2018), "Never-ending learning", *Communications of the ACM*, Vol. 61/5, pp. 103-155, http://dx.doi.org/10.1145/3191513. [3]

Rapp, D. and J. Braasch (eds.) (2014), *Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences*, MIT Press, Cambridge, MA. [11]

Rips, L. (1994), *The Psychology of Proof: Deduction in Human Thinking*, MIT Press, Cambridge, MA. [13]

Sinatra, A. et al. (2021), *Design Recommendations for Intelligent Tutoring Systems: Competency-based Scenario Design*, Army Research Laboratory, Orlando, FL. [9]