# 11. Identifying artificial intelligence capabilities: What and how to test

José Hernández-Orallo, Universitat Politècnica de València

Evaluating the capabilities of artificial intelligence (AI) has enormous implications in many areas, especially for the future of work and education. The context is also changing rapidly: the capabilities of humans and AI co-evolve, with scenarios of replacement, displacement or enhancement. Beginning with a review of several taxonomies from human evaluation and AI, this chapter presents a 14-ability taxonomy to identify abilities as potentially disassociated clusters to characterise AI systems. It explores a range of human tests used for decades in recruitment and education, contrasting them with the increasing trend towards basing AI evaluation on benchmarks. The chapter reviews the challenges of bringing human tests to evaluate AI, identifying guidelines to devise reliable tests to compare the capabilities of humans and AI.

## Introduction

This chapter analyses how the evaluation of artificial evaluation (AI) systems differs from that of other hardware and software systems, and how it diverges from the evaluation of human cognition – from abilities to skills. It covers common skill taxonomies used for humans as opposed to those used in subfields in AI. However, it proposes a taxonomy based on abilities since skills, knowledge and task performance, if not programmed specifically, must ultimately develop from abilities.[1]

This chapter discusses a recently introduced taxonomy, including 14 abilities, that could be useful for both humans and AI. It examines measurement problems of tests for human evaluation (psychometric, educational and professional) and the bevy of AI evaluation platforms. Subsequently, it argues that human tests cannot be directly used as measurement instruments for an ability-oriented evaluation of AI.

Nonetheless, it identifies the elements of tests that should be abandoned and those that could be reused for working adaptations or newly created tests. It also outlines pragmatic solutions of mapping human tests and AI benchmarks through the ability taxonomy. It ends with recommendations to evaluate AI systems more precisely to determine what they can really do.

## Future artificial intelligence roles

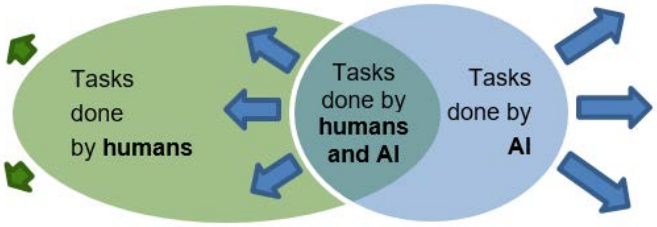
### *Displacement vs. replacement*

Some narratives envision that AI will "replace" humans[2], taking over highly skilled and enjoyable activities from engineering to the arts. However, with AI being around for a long time, history suggests that AI will first transform work, creating new kinds of occupations and tasks for humans to do on their own and to share with machines. For instance, the portfolio of a handful of clients handled by a human has been transformed into a more sophisticated framework dealing with massive portfolios of clients. Such transformation was made possible by machine learning models and other AI techniques such as planning, optimisation, natural language processing and sentiment analysis. These AI components now talk with customers, anticipate their behaviour, promote cross-selling and even ensure their overall satisfaction with the company.

Even in cases where the whole task is seemingly automated, a deeper analysis shows that humans usually assume subsidiary new subtasks so the system can work. This phenomenon is not new, as when repurposed workers had to oil or repair a newly introduced machine. However, this kind of "incomplete automation" is easier to hide with cognitive tasks, which humans now do (even unpaid, such as when a customer orders with their phone at a "restaurant").

Variants of this phenomenon have been referred to as *fauxtomation* (Jackson, 2019[1]) or simply "human computation" (Von Ahn, 2008[2]; Taylor, 2018[3]). Figure 11.1 shows how the narrative of AI increasingly taking on more tasks than humans today is incomplete. While some new tasks are created for and only done by machines, humans are doing new tasks, too.

Figure 11.1. New tasks are done by humans, by AI and by both together




In this transformation, there is displacement rather than just replacement. Human labour has not disappeared in areas where AI has had an important impact. Humans do different tasks, and many of them are new. Typically, these tasks relate to setting goals and targets, monitoring robots and other AI systems, adapting and integrating their decisions, and curating "training" material for AI systems.

This narrative of displacement vs. replacement will determine the path of AI research. The prevalence of responsible and human-centred AI today is introducing a culture based on three factors. First, humans must always be in the loop. Second, humans and machines should collaborate and not compete. Third, machines should always be subordinate to humans.

This vision suggests the separation between what AI will be able to do from what it will be allowed to do. For instance, one day it will be possible to automate most of a physician's tasks. However, this may never happen because of ethical issues or mainstream social pressure. This is not new or particular to AI; the "human touch" is considered a special value in some domains that can already be automated (e.g. hand-made delicatessen). However, ethical and political decisions more than technology may dominate decisions about what AI can do and what is *reserved* for humans in the future.

### *Human-machine collaboration through externalisation and extension*

Another important factor is the different forms in which humans and machine collaborate, or create new behaviour (Rahwan, 2019[4]). In basic "externalisation", a human delegates a task to a machine, giving it the instructions, goals or input. However, humans are also integrating AI capabilities in a more coupled way through "AI extenders" (Hernández-Orallo and Vold, 2019[5]). For instance, most humans use navigational tools in their phones to go to a new address. While the tool shows the position and optimal routes, the human still navigates the city and ultimately decides where to go and how. In this way, the full cognitive process of going from position A to B is not externalised but extended.

In general**,** machines complement or extend humans, rather than replace them. Humans then adapt to the new situation, developing new "digital" skills (e.g. using new AI apps). This means that human skills are changing significantly. Young people text extremely fast because they use the predictive hints given by their instant messaging app.

This phenomenon – which goes beyond skills – is changing abilities as well. For instance, factual memory capabilities are falling because people can check any fact on the Internet. This is known as the "Google effect" (Bohannon, 2011[6]). Provided the technology does not fail, the new situation (and the associated atrophy) should be seen as empowering. When coupled with their AI extenders, humans should thus be considered more capable overall.

Accordingly, as humans can do more things using tools, they should be evaluated with those tools. As calculators are allowed in many math exams, AI extenders should be allowed for writing, editing, drawing or speaking in more proficient and creative ways. Almost no one writes today without spell check or online access to Wikipedia. Limiting access to AI tools and extenders in tests would thus measure an unrealistic situation. The implications of all this for the following sections are important:

- Tasks humans do are changing more rapidly.
- Human skills, and even some specific abilities, are changing.
- Human evaluation tests are (or should be) changing.

In a rapidly changing world fuelled by AI, tasks are going to change faster, and so will skills and knowledge. This affects humans but also AI. For AI to become economically advantageous over human labour, it needs general abilities rather than specialised skills. In this way, AI systems can learn new skills efficiently and autonomously, adapting faster than humans to new tasks and procedures.

## From skill lists to ability taxonomies

### *Evaluating artificial intelligence through its abilities rather than task completion*

Evaluating what AI does in terms of tasks would be short-sighted, unlikely to be comprehensive and prone to overfitting.[3] An AI or robotic system showing performance at a particular task gives poor indications of what other things AI can do. For instance, while computer chess reached superhuman level in the late 1990s (Weber, 1997[7]; Campbell, Hoane and Hsu, 2002[8]), it took AI 20 years to reach human-level performance on other "similar" games such as Go (Silver et al., 2016[9]).

Quite remarkably, both the chess and Go milestones used completely different approaches. Similarly, floor robotic cleaners have been in residential homes for quite some time, but no robot can yet clean a table full of objects properly. For these and other reasons, an ability-oriented[4] evaluation in AI should be preferred over a task-oriented evaluation (Hernández-Orallo, 2017[10]; Hernández-Orallo, 2017[11]).

While still unusual in AI, the standard approach to evaluate humans is by abilities and skills rather than through tasks. Specific tasks for a job, such as driving a lorry, are evaluated only occasionally. Moreover, this evaluation is usually accompanied with verification of some other skills, knowledge and abilities, while many others are taken for granted (e.g. being able to understand an order from the boss).

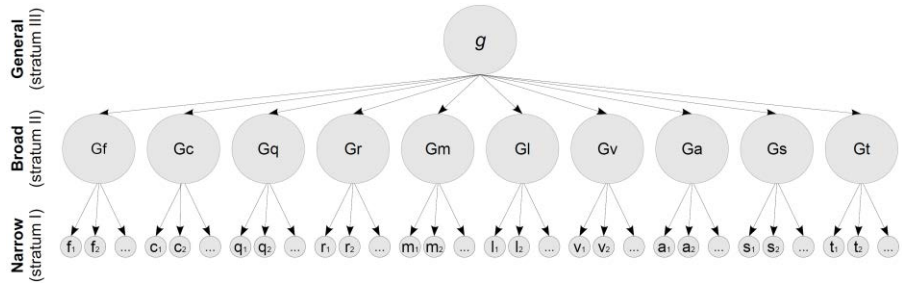
### *Defining abilities and skills*

Before looking at taxonomies of abilities and skills, the terms should be clarified. Abilities and skills are sometimes used interchangeably, but skills and knowledge are more properly used for effective competences. For example, solving differential equations, programming in Python, playing the guitar or working in teams is a skill. Abilities usually refer to potential capacities related, for example, to verbal, spatial or metacognitive domains.

Abilities allow humans to acquire and deploy skills and knowledge (Schuelke and Day, 2012[12]). This interpretation goes beyond clustering tasks into skills, and skills into abilities. Rather, it means that abilities allow individuals to develop skills and knowledge, which leads to further skills and, ultimately, the capacity to do tasks. Also, abilities in humans are stable for a significant part of the adult life. Conversely, skills are learnt, improved or even forgotten at different moments in the life of an individual.

There are many taxonomies for human cognitive abilities, and some differ significantly. The Cattell-Horn-Carroll (CHC) taxonomy (Carroll, 1993[13]) characterises humans hierarchically (Figure 11.2).

## Figure 11.2. Carroll's three-stratum model




Note: This model is usually known as Cattell-Horn-Carroll taxonomy. The g factor is on the top (third) level, ten broad abilities are at the second level, with the bottom (first) level including many more "narrow" abilities.

### *The Cattell-Horn-Carroll taxonomy*

Table 11.1 describes the ten factors in stratum II of the CHC taxonomy. Some, like processing speed, are not really "abilities". Rather, they are "factors" identified through techniques such as factor analysis. Their interpretation requires examination of their loadings, connecting with underlying tests and the lower stratum.

Nonetheless, these factors capture some areas of cognition and intelligence for which humans involve different intrinsic mechanisms. For example, inductive inference is associated with Gf and Gc. Meanwhile, deductive inference is associated with Gq or neurological substrate (e.g. Gv and Ga).

## Table 11.1. Broad abilities at stratum II of Cattell-Horn-Carroll taxonomy

| Factor | Ability |
|---|---|
| Gf | Fluid intelligence |
| Gc | Crystallised intelligence |
| Gq | Quantitative reasoning |
| Gr | Reading and writing ability |
| Gm | Short-term memory |
| Gl | Long-term storage and retrieval |
| Gv | Visual processing |
| Ga | Auditory processing |
| Gs | Processing speed |
| Gt | Reaction time/decision speed |

### *Developmental perspective*

Cognitive development offers a more dynamic take on what an individual can do. Piaget introduced four main stages in child development (Piaget, 1936[14]; Piaget, 1964[15]): sensorimotor (0-2 years), preoperational (2-7 years), concrete-operational (7-12 years) and formal-operational (12 and more years). Many variants and extensions have followed, sometimes referred to as neoPiagetian models (Morra et al., 2012[16]).

The developmental perspective studies how skills are built over other skills. From the perspective of education, the focus is on skills, as many can be acquired in some degree at any age. Conversely, abilities have a greater innate component and consolidate in adolescence. However, ignoring abilities altogether leads to poor understanding of why some individuals develop some skills better than others.

### *Taxonomies for job analysis*

The use of skills and knowledge is more common in the analysis of the workplace. Taxonomies for job analysis such as O*NET-SOC, ISCO and ESCO. O*NET-SOC[5] includes more than 1 000 "occupational titles" classified into 23 major groups[6], which are mostly sector-oriented (e.g. health, construction, etc.). ISCO[7] includes more than 1 500 categories of occupations, organised into ten major groups (Table 11.2). ESCO[8] covers about 3 000 occupations and around 13 500 skills. It is partially hierarchical. Apart from a sector-oriented hierarchy, it has a top hierarchy of skills that is similar but different from ISCO, including eight categories (Table 11.3).

**Table 11.2. Ten major groups in the ISCO occupation categories**

| Code | Category |
|---|---|
| 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 and related trades workers |
| 8 | Plant and machine operators, and assemblers |
| 9 | Elementary occupations |
| 0 | Armed forces occupations |

**Table 11.3. Eight major groups in the ESCO occupational skills**

| Code | Category |
|---|---|
| S1 | Communication, collaboration and creativity |
| S2 | Information skills |
| S3 | Assisting and caring |
| S4 | Management skills |
| S5 | Working with computers |
| S6 | Handling and moving |
| S7 | Constructing |
| S8 | Working with machinery and specialised equipment |

The cognitive ability taxonomies, developmental models and job competence classifications are derived from and aimed at humans. When taxonomies are derived from human populations, some elements that are essentially different may fall into the same human ability, simply because they are correlated in the human species. As well, similar elements may be separated because they are handled by different modules or genes in humans.

In cognitive development, some stages seem more universal than others, but even for some animals the stages may differ significantly. Foals and other precocial animals can stand and run in a few hours, for example. Similarly, the competences and skills used for labour depend on the society and the economy at a particular time and culture. Neither of these human taxonomies is immediately applicable to AI.

Domains in AI are strongly linked to underlying techniques, which have varied significantly in a few decades. AI typically organises its functionalities with terms such as learning, planning, recognition, inference, etc. To date, there is no standard taxonomy of AI skills or abilities. Areas and domains are typically used for textbooks and conferences or bibliometric research (Machado et al., 2018[17]; Frank, Wang and Cebrian, 2019[18]).

### *A taxonomy of cognitive abilities*

Taking all this into account, Hernández-Orallo and Vold (2019[5]) introduce a taxonomy of cognitive abilities, merging several categorisations in psychology, animal cognition and AI. To be comprehensive about all cognitive abilities, the methodology started with elements from all these disciplines, distilling from different sources:

- Thurstone's primary mental abilities according to factors from CHC hierarchical model (stratum II, Figure 11.2)
- areas of animal-cognition research according to Wasserman and Zental (2006[19])
- main areas in AI according to the *AI Journal* (as per 2017)
- "competency" areas in AGI according to Adams et al. (2012[20])
- I-athlon "events" from Adams, Banavar and Campbell (2016[21]).[9]

These different lists were integrated by matching synonyms and related terms, and trying to keep a manageable number of broad capabilities. There were tensions between both distinctiveness and comprehensiveness against the number of abilities. The main criterion for distinguishing between two abilities A and B (and not merging them) was the understanding that a system or component (either natural or artificial) could *conceivably* master one of them while failing at the other. The compromise for completeness was easier to find. Some elements (such as processing or decision speed in the CHC) are not proper abilities. In addition, some abilities related to multimodality were not explicitly included in the final list of 14 (e.g. olfactory processing). The current version only covers "visual" and "auditory" processing, being the two most representative sensory modalities.

The taxonomy is further developed into a rubric in Martínez-Plumed et al. (2020[22]). The 14 cognitive abilities are shown in Table 11.4 and a number of principles behind their development are presented in Box 11.1.

## Table 11.4. Cognitive abilities applicable to both humans and AI systems

| Ability | Description |
|---|---|
| MP: Memory processes | Storage of information in an appropriate medium to be recovered at will according to some keys, queries or mnemonics. This covers long-term memory and episodic memory. |
| SI: Sensorimotor interaction | Perception of things, recognising patterns and manipulating them in physical or virtual environments with parts of the body (limbs) or other actuators, through various sensory and actuator modalities, and representations. |
| VP: Visual processing | Processing of visual information, recognising objects and symbols in images and videos, movement and content in the image, with robustness to noise and different angles and transformations. |
| AP: Auditory processing | Processing of auditory information, such as speech and music, in noisy environments and at different frequencies. |
| AS: Attention and search | Focusing attention on the relevant parts of a stream of information in any kind of modality, by ignoring irrelevant objects, parts, patterns, etc. Similarly, seeking those elements that meet some criteria in the incoming information. |
| PA: Planning, sequential decision making and acting | Anticipating the consequences of actions, understanding causality and calculating the best course of actions given a situation. |
| CE: Comprehension and compositional expression | Understanding natural language, other kinds of semantic representations in different modalities, extracting or summarising their meaning, as well as generating and expressing ideas, stories and positions. |
| CO: Communication | Exchanging information with peers, understanding what the content of the message needs for a given effect, following different protocols and channels of informal and formal communication. |
| EC: Emotion and self-control | Understanding the emotions of other agents, how they affect their behaviour and also recognising their own emotions and controlling them and other basic impulses depending on the situation. |
| NV: Navigation | Moving objects or oneself between different positions, through appropriate, safe routes and in the presence of other objects or agents, and changes in the routes. |
| CL: Conceptualisation, learning and abstraction | Generalising from examples, receiving instructions, learning from demonstrations and accumulating knowledge at different levels of abstraction. |
| QL: Quantitative and logical reasoning | Representation of quantitative or logical information that is intrinsic to the task, and the inference of new information from them that solves the task, including probabilities, counterfactuals and other kinds of analytical reasoning. |
| MS: Mind modelling and social interaction | Creation of models of other agents to understand their beliefs, desires and intentions, and anticipate the actions and interests of other agents. |
| MC: Metacognition and confidence assessment | Evaluation of their own capabilities, reliability and limitations, self-assessing the probability of success, the effort and risks of own actions. |

Source: Hernández-Orallo and Vold (2019[5]).

### Box 11.1. Principles behind the cognitive ability taxonomy

There are some principles behind the taxonomy in Table 11.4:

- First, the clusters should not be informed by the categories in human or AI taxonomies only. Abilities should be identified as different when, with the current knowledge, they are thought to conceivably rely on different mechanisms (e.g. deductive and inductive inference).
- Second, it is convenient to associate taxonomies with rubrics that determine whether a task or skill requires the ability. This can be understood as a representational definition and understanding for each ability, and not simply as a cryptical latent "factor" or a meaningless construct.
- Third, developing a test that only measures one ability for every kind of subject (natural or artificial) is complicated. It is more practical to think of many-to-many quantitative connections between abilities and tests, with the advantage of reusing the results of existing tests and benchmarks.
- Finally, skills must always be connected with abilities in the context of development. For instance, the progression of an AI system in the acquisition of elemental skills can be a good way to ensure the system has the abilities needed to develop the skills.

Source: Hernández-Orallo and Vold (2019[5]).

The above taxonomy is not static. However, it serves as a stable source to do mappings between other AI/human taxonomies (Martínez-Plumed et al., 2020[22]; Tolan et al., 2021[23]) and especially tests, as explored in the following section.

## Tests: Caveats and pathways

### *Evaluating humans*

Abilities[10] are usually latent variables or constructs, with tests being instruments for measuring them. During the 20th century, a plethora of tests was developed for evaluating humans:[11]

- Psychometric tests for general abilities

These notably include those related to IQ tests. Example: Wechsler Adult Intelligence Scale, with tests aggregated into four categories: verbal comprehension, perceptual reasoning, working memory and processing speed.[12]

- Developmental tests

These cover a series of stages for different purposes (e.g. detecting disabilities). Example: the Bayley scales (Bayley, 1993[24]) evaluate children from ages to 3.5 years with items in three categories: mental scale, motor scale and behaviour rating scale.

- Tests for consolidated knowledge or general education skills

These explore "attainment" or "achievement". For instance, military psychometric tests (such as Armed Services Vocational Aptitude Battery) and college entrance exams (such as ACT and SAT) cover a mixture of abilities and skills (English, mathematics, reading, writing and science). The Bennett Mechanical Comprehension Test covers more specific abilities and skills.

- Personnel selection and certificates

This combines psychometric tests, interviews and practical demonstrations to certify certain abilities, attitudes, knowledge and skills. In many countries, for instance, a driver's licence test evaluates reaction time, visual acuity, knowledge and ability to judge traffic signs and rules. It combines these tests with a practical exam with a real car and sometimes a simulator.

### *Evaluating machines*

Given the range and diversity of evaluations for humans, what can be used to evaluate machines? AI evaluation differs from the evaluation of many other software and hardware systems. For AI, there is usually no formal or procedural description of how the system must solve a goal (otherwise, the solutions would be programmed). Experimental evaluation then becomes more relevant in AI (from learning to planning) than in other areas of computer science.

Apart from informal and subjective assessments (e.g. the Turing test), AI has rubrics and benchmarks.

*Rubrics* are generally based on human assessment about the capability of the system. Unlike open evaluations such as the Turing test, rubrics are systematic. For instance, (Brynjolffson and Mitchell, 2017[25]; Brynjolffson, Mitchell and Rock, 2018[26]) present a series of questions about a task (the rubric), giving a score that represents whether machine learning could automate the task. A recent OECD project (Elliott, 2017[27]) relies on subject matter experts to assess AI capabilities for three areas in the Programme for the International Assessment of Adult Competencies. A more general approach is based on technology readiness levels. Here, the rubric distinguishes different levels, from research ideas in the lab to viable products (Martínez-Plumed, Gómez and Hernández-Orallo, 2020[28]; Martínez-Plumed, Gómez and Hernández-Orallo, 2021[29]).

*Benchmarks* are repositories of instances of a task (or collections of tasks) that serve as challenges for AI to improve on several metrics of *performance*. Benchmarks are undoubtedly fuelling the progress of the field (Hernández-Orallo et al., 2016[30]) but are still limited as valid measurement instruments. AI systems commonly reach superhuman performance on a benchmark but do not display the associated capability; the systems usually fail beyond the conditions and distribution of the benchmarks. Accordingly, many benchmarks are soon replaced, entering a "challenge-solve-and-replace" (Schlangen, 2019[31]) or a "dataset-solve-and-patch" (Zellers et al., 2019[32]) dynamic.[13]

### *Assessing artificial intelligence with tests designed for humans*

Given these validity problems in AI evaluation, and the wide range of valid tests for humans, using tests designed for humans might seem a good idea for AI. There are several reasons why this is not advisable:

1. Tests are devised as measuring instruments for a particular population. Human tests lack measurement invariance beyond the human population (even beyond adults).
2. Humans are embodied agents. Many AI systems do not take the form of an agent, and sometimes not even the form of a system. Instead, they appear as cognitive components or modules.
3. A single human can perform well for many tasks and tests. When AI is said to solve A and B, for example, this typically means that one AI system solves A and another AI system solves B.
4. AI systems and components can be built on purpose for a task. The designers put a lot of specific knowledge, bias or curated training data for the particular benchmark.
5. The behavioural traits of humans and AI overlap. AI may "conquer" more human abilities in the future, but AI is introducing many other new abilities (see Figure 11.1).
6. Humans and AI differ on the resources used (e.g. data, compute, sensors) or ignore/ban associated human cognitive labour (e.g. labelling data, delegation to human computation).

The six reasons are exemplified by the use of IQ and other human intelligence tests as benchmarks for AI. Whenever a type of IQ problem or a battery of intelligence tests is made available for AI researchers, less and less time, but increasingly more computational resources are needed for a new AI system to excel at the tests (Hernández-Orallo et al., 2016[33]). However, this AI system can do nothing else beyond the particular IQ tests. Remarkably, such tests are not about knowledge and specialised skills but rather about *human* core reasoning capabilities and abstract problems. They include letter series, number series, Raven's progressive matrices, odd-one-out problems, vocabulary analogies, geometric analogies, etc. The success of AI systems on these intelligence tests has not shown real progress in AI. It cannot be used as evidence that AI systems have general intelligence (Hernández-Orallo et al., 2016[30]).

While intelligence tests are just a kind of cognitive test for humans, other human tests are also problematic when applied to AI (Dowe and Hernández-Orallo, 2012[34]; Hernández-Orallo, 2017[11]). For instance, AI challenges have used questions from educational exams, including diagrams, geometry and mathematics from 4th grade science exams (Clark and Etzioni, 2016[35]). The bad results were reassuring: "no system [came] even close to passing a full 4th grade science exam". Good results from the Aristo Project and other (ensembles of) language models just three years later were labelled as "a significant milestone" (Clark et al., 2019[36]).

However, the new AI solutions are not really "general question-answering" systems and cannot compare to a human with a similar result. Massive language models certainly solve many questions – more than the average human – but a closer look reveals the system is not robust to minor variations of the questions. Ultimately, it does not really understand the questions. In other words, the positive test results for AI on human tests, when compared to humans, are hugely overestimated because of overfitting.

### *More promising avenues*

To avoid the recurrent problem of overfitting, some new benchmarks for AI are taking inspiration from human tests but have been profoundly modified or reconstructed. The tests aim to be easy for humans but challenging for state-of-the-art AI. However, they should not contain hidden statistical patterns or other artefacts that AI systems could exploit to circumvent what the inspirational tests are supposed to measure in humans.

#### *Winograd and Winogrande*

The Winograd Schema Challenge has been one of the most important attempts in this direction. It was presented as a collection of text comprehension questions using pronouns that must be disambiguated (Levesque, Davis and Morgenstern, 2012[37]). Levesque initially sought to design questions whose answer would show a high level of common sense reasoning around the elements appearing in the question. However, several AI systems have recently shown excellent performance by exploiting some statistical artefacts in the way the questions are generated. These systems use "clever tricks involving word order or other features of words or groups of words". However, they do not really display the capabilities for referential disambiguation that the test is assumed to be measuring (Kocijan et al., 2020[38]).

Winogrande is a much larger version meant to replace Winograd's schemas. However, new language models have quickly reached good performance too, while still being far from general language understanding. This happens in all areas of AI, from natural language to machine vision. It is sometimes called the Clever Hans phenomenon, as AI finds alternative cues and tricks to solve the task in the same way as a celebrated 19th century horse did to amaze spectators (Lapuschkin et al., 2019[39]; Hernández-Orallo, 2019[40]). Due to the Clever Hans phenomenon, the validity of many tests for AI systems is constantly questioned. This, in turn, provokes the "challenge-solve-and-replace" (Schlangen, 2019[31]) dynamics mentioned earlier. At its heart it reveals an *adversarial* game between AI developers (and their systems) and the evaluators (Hernández-Orallo, 2020[41]). Such an adversarial philosophy is intrinsic to evaluation and should be incorporated in the design of benchmarks and evaluation procedures.

#### *New benchmarks from natural language processing*

There are several good examples of these new benchmarks in natural language processing.[14] MOSAIC, for example, includes the adversarial generation of examples found in SWAG (Zellers et al., 2018[42]) or DynaBench.[15] An adversarial example is modified slightly such that humans are not significantly affected, while AI systems fail catastrophically. Understanding how these examples must be generated for different kinds of AI systems can help improve the systems.

Ultimately, if the only possible examples that make an AI system fail also make humans fail, the AI system may really be better than humans. At this point, the question of what the test measures, and all the variations of the instances that become part of the measure, can be considered. As AI systems become designed for the test, an adversarial mindset is needed more than for the evaluation of humans. Training to the test also happens for humans but to a lesser extent.

#### *Non-human animal tests and sandbox evaluation*

A less anthropocentric stance to the evaluation of AI looks at non-human animals. In the 1990s, for instance, the Cognitive Decathlon was built for DARPA's Biologically Inspired Cognitive Architecture programme, based on developmental tests (Mueller, 2010[43]). The battery was discontinued around ten years ago (Mueller et al., 2007[44]). However, in a related approach, the animal-AI environment builds on animal tests rather than human tests (Crosby et al., 2019[45]; Crosby et al., 2020[46]). If the adaptation of these tests become more common in the future, the main risk would still be overfitting the AI system to the test distribution, instead of really solving the constructs the test was supposed to measure.

Sandbox evaluation provides a possible solution to overfitting. In these cases, rather than training instances, AI developers are provided with a “sandbox environment” to create different curricula for the AI system. Only when the system has been “raised” in the environment can evaluators disclose the tasks, and test the system, without further training or adaptation. The idea is to encourage the construction of systems that can master a domain, rather than mastering tasks from a distribution. The Animal-AI Olympics followed this philosophy (Crosby et al., 2019[45]; Crosby et al., 2020[46]).

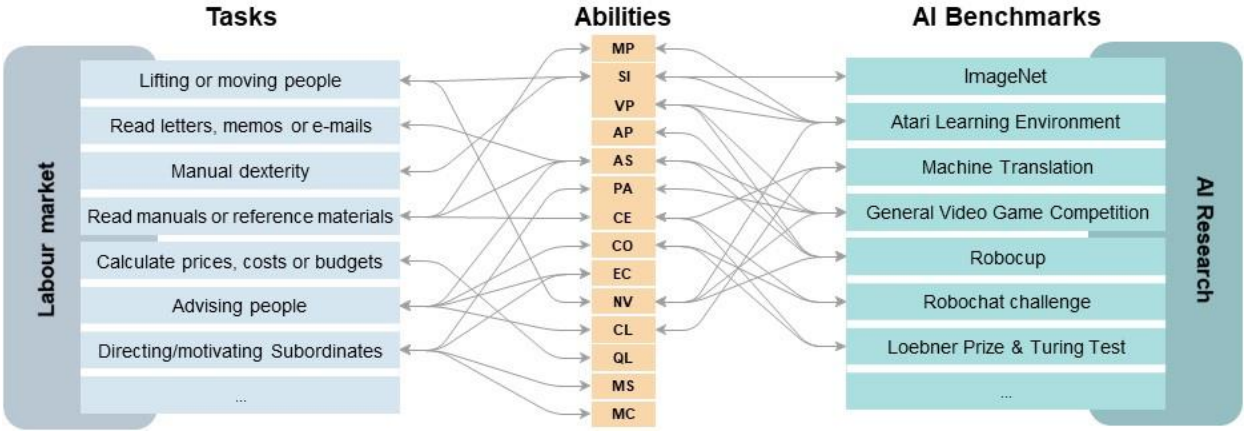
### *Expert judgement*

More immediate options to compare human and AI capabilities are needed until new tests inspired by human or animal evaluation or based on more fundamental principles are developed (Hernández-Orallo, 2017[11]). A pragmatic alternative to tests is the use of human experts to assess AI capabilities through rubrics or other kind of questionnaires.

#### *Indirect mapping between job market and AI benchmarks*

Martínez-Plumed et al. (2020[22]) explore a hybrid solution that performs a mapping between AI benchmarks and the 14 cognitive abilities in Table 11.4. Through expert questionnaires and other methodologies (e.g. Delphi), a matrix of many-to-many correspondences is established between benchmarks and abilities. As the performance for different AI benchmarks is usually incommensurate, Martínez-Plumed et al. measure the intensity of research in terms of number of related papers or media articles. This measure, in turn, is mapped to intermediate abilities. By doing a similar mapping from abilities to other elements, such as occupational tasks, AI benchmarks can be linked with occupations (Figure 11.3).

## Figure 11.3. Bidirectional and indirect mapping between job market (ISCO-3 specifications) and AI benchmarks




Source: (Martínez-Plumed et al., 2020[22]).

Bidirectional and indirect mapping is a promising approach but must be used cautiously for quantifying abilities. The "latent" intermediate abilities could, in principle, be mapped to other elements, such as human test results. However, the performance results of different benchmarks should not be aggregated; research intensities can be calculated for the ability but not the magnitude of the ability. This is because results at the leader boards for all benchmarks are about different AI systems. The systems that "solve" ImageNet are different from those used for Robocup.

More importantly, even if the same general AI system is used for many of the benchmarks (e.g. a language model), the magnitudes of performance (the scales) are different (Hernández-Orallo and Vold, 2019[5]; Hernández-Orallo, 2020[41]). Indeed, 90% success in ImageNet cannot be averaged with 72% correct answers in Winogrande or a score of 570 points in Pacman.

Normalising results against a human population is possible. However, the transformation should be based on *percentiles* over the human population rather than using human average performance. Moreover, this does not solve the scaling problem, as the relevance of each test would depend on the variance of the human population for that test.

### *AI Collaboratory*

Some initiatives are exploring better mappings and aggregations of evaluation results for AI and humans to allow for meaningful comparison. One such initiative is the AI Collaboratory (Martínez-Plumed, Gómez and Hernández-Orallo, 2020[47]; Martínez-Plumed, Gómez and Hernández-Orallo, 2020[48]; Martínez-Plumed, Gómez and Hernández-Orallo, 2020[49]). As part of the AI WATCH programme (Martínez-Plumed, Gómez and Hernández-Orallo, 2020[47]) of the European Commission, it collects and structures evaluation results for AI and humans, and building mappings and hierarchies.

The AI Collaboratory is structured with a multidimensional schema. It contains information about the facts (the measurements) and satellite information about *who* is measured (the intelligent systems), *what* is measured (the tests) and *how* it is measured (the procedures). Each dimension is hierarchical. For instance, in the "who" dimension, systems can be aggregated into populations, populations into families, etc. In the "what" dimension, examples can be aggregated into tests, tests into batteries, etc. A taxonomy of abilities, as those seen in Figure 11.2 or Table 11.4, could be easily defined in the "what" dimension. Despite all the caveats for comparing human results with AI results, data-driven tools and meta-analyses are essential to understand how measurements relate to each other.

## Recommendations

With the dominance of the machine learning paradigm, and skills changing more rapidly, AI systems will likely become less specialised for particular skills and tasks to be profitable. Consequently, a taxonomy of abilities, such as the one shown in Table 11.4, using the principles in Box 11.1, serves as a foundation for the evaluation of more adaptable AI systems. There will be exceptions like testing standardised skills or tasks, such as driving, but more general, ability-oriented AI will be able to adapt to evolving tasks and skills required at home and in the workplace.

- **Develop new testing protocols and share detailed results for data-driven exploration**

It is not enough to think in terms of abilities rather than tasks, or to build new benchmarks that cover an ability rather than a task. Instead, new testing protocols in AI are needed that go beyond training-test, avoiding learning specialisation or even AI systems built for the test. Also, evaluations should consider all the resources and costs involved in the solution (data, compute, human computation, etc.) (Martínez-Plumed et al., 2018[50]). Evaluation must become more iterative, more adversarial to avoid being gamed by AI researchers (willingly or not) (Hernández-Orallo, 2020[41]). Finally, there is a lack of meta-analysis and data-driven exploration of AI capabilities.

- **Use an intrinsic scale for new test designs**

Lack of common categories, and especially of commensurate scales, is a critical concern when reusing results from different AI benchmarks, and especially when comparing them with human tests results. New test designs should use an intrinsic scale, independent of the population to be tested. As an ultimate resource, scores could be normalised according to the distribution of human results rather than as a single individual or population average. AI often sets this average as a misleading threshold for "human-level performance".

- **Learn from human evaluation**

Use of human tests for evaluating AI *directly* is not feasible for a number of reasons. However, ignoring human evaluation, its tests and its associated techniques would be a mistake. Such approaches offer lessons to learn from. They may imply a cognitive overhaul of evaluation in AI. Exploring mappings and meaningful aggregations that capitalise on the information from human tests and AI benchmarks is a worthwhile initiative. AI and robotics – and human hybridisations yet to come – deserve more than simple performance-based, task-oriented evaluation.

## References

Adams, S. et al. (2012), "Mapping the landscape of human-level artificial general intelligence", *AI Magazine*, Vol. 33/1, pp. 25-42. [20]

Adams, S., G. Banavar and M. Campbell (2016), "I-athlon: Towards a multidimensional Turing test", *AI Magazine*, Vol. 33/1, pp. 25-42. [21]

Bohannon, J. (2011), "Searching for the Google effect on people's memory", *Science*, Vol. 335/15 July, p. 277. [6]

Brynjolffson, E. and T. Mitchell (2017), "What can machine learning do? Workforce implications", *Science*, Vol. 358/6370, pp. 1530-1534. [25]

Brynjolffson, E., T. Mitchell and D. Rock (2018), "What can machines learn and what does it mean for occupations and the economy?", *AEA Papers and Proceedings*, Vol. 108/May, pp. 43-47. [26]

Campbell, M., A. Hoane and F. Hsu (2002), "Deep blue", *Artificial Intelligence*, Vol. 134, pp. 55-83. [8]

Carroll, J. (1993), *Human Cognitive Abilities: A Survey of Factor-analytic Studies*, Cambridge University Press, New York. [13]

Clark, P. and O. Etzioni (2016), "My computer is an honor student–But how intelligent is it? Standardized tests as a meausre of AI", *AI Magazine*, Vol. 37/1, pp. 5-12. [35]

Clark, P. et al. (2019), "From 'F' to 'A' on the N.Y. Regents Science Exams: An overview of the Aristo Project", *arXiv*, Vol. 1909.10958. [36]

Crosby, M. et al. (2019), "Translating from animal cognition to AI", *NeurIPS 2019 Competition and Demonstration Track, PMLR*, pp. 164-176. [45]

Crosby, M. et al. (2020), "The animal-AI testbed and competition", *NeurIPS 2019 Competititon and Demonstration Track, PLMR*, pp. 166-176. [46]

Dowe, D. and J. Hernández-Orallo (2012), "IQ tests are not for machines, yet", *Intelligence*, Vol. 40/2, pp. 77-81. [34]

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. [27]

Firestone, C. (2020), "Performance vs. competence in human-machine comparisons", *Proceedings of the National Academy of Sciences*, Vol. 117/43, pp. 26562-26571. [51]

Fleishman, E. (1972), "On the relation between abilities, learning and human performance", *The American Psychologist*, Vol. 27/11, pp. 1017-1032. [52]

Frank, M., D. Wang and M. Cebrian (2019), "The evolution of citation graphs in artificial intelligence research", *Nature Machine Intelligence*, Vol. 1/2, pp. 79-85. [18]

Hamilton, E., J. Rosenberg and M. Akcaoglu (2016), "The Substitution Augmentation Modification Redefinition (SAMR) model: A critical review and suggestions for its use", *TechTrends*, Vol. 60, pp. 433-441. [53]

Hernández-Orallo, J. (2020), "Twenty years beyond the Turing test: Moving beyond the human judges too", *Minds & Machines*, Vol. 30, pp. 533-562. [41]

Hernández-Orallo, J. (2019), "Gazing into Clever Hans machines", *Nature Machine Intelligence*, Vol. 1/4, pp. 172-174. [40]

Hernández-Orallo, J. (2017), "Evaluation in artificial intelligence: From task-oriented to ability-oriented measurement", *Artificial Intelligence Review*, Vol. 48/3, pp. 398-447. [10]

Hernández-Orallo, J. (2017), *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*, Cambridge University Press, New York. [11]

Hernández-Orallo, J. et al. (2016), "A new AI evaluation cosmos: Ready to play the game", *AI Magazine*, Vol. 38/3, pp. 66-69. [30]

Hernández-Orallo, J. et al. (2016), "Computer models solving intelligence test problems: Progress and implications", *Artificial Intelligence*, Vol. 230, pp. 74-107. [33]

Hernández-Orallo, J. and K. Vold (2019), "AI extenders: The ethical and societal implications of humans cognitively extended by AI", *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Association for Computing Machinery, New York. [5]

Jackson, G. (2019), "Why the rise of the robots hasn't happened just yet", 23 January, The Financial Times, https://www.ft.com/content/ec2f65c8-1e61-11e9-b2f7-97e4dbd3580d. [1]

Kocijan, V. et al. (2020), "A review of Winograd schema challenge datasets and approaches", *arXiv preprint arXiv:2004*, Vol. 13831. [38]

Krizhevsky, A. (2009), *Learning Multiple Layers of Features from Tiny Images*, University of Toronto, https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. [54]

Lapuschkin, S. et al. (2019), "Unmasking Clever Hans predictors and assessing what machines really learn", *Nature Communications*, Vol. 10/1, pp. 1-8. [39]

Levesque, H., E. Davis and L. Morgenstern (2012), "The Winograd schema challenge", presentation, Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning. [37]

Machado, M. et al. (2018), "Revisting the arcade learning environment: Evaluation protocols and open problems for general agents", *Journal of Artificial Intelligence Research*, Vol. 61, pp. 523-562. [17]

Martínez-Plumed, F. et al. (2018), "Accounting for the neglected dimensions of AI progress", *arXiv preprint arXiv*, Vol. 1806.00610. [50]

Martínez-Plumed, F., E. Gómez and J. Hernández-Orallo (2021), "Futures of artificial intelligence through technology readiness levels", *Telematics & Informatics*, Vol. 58/101525. [29]

Martínez-Plumed, F., E. Gómez and J. Hernández-Orallo (2020), "AI Watch: Methodology to monitor the evaluation of AI technologies", *JRC Working Papers*, No. 120090, Joint Research Centre. [49]

Martínez-Plumed, F., E. Gómez and J. Hernández-Orallo (2020), *Assessing Technology Readiness Levels of Artificial Intelligence*, Joint Research Centre Report, AI Watch, European Commission, Brussels. [28]

Martínez-Plumed, F., E. Gómez and J. Hernández-Orallo (2020), "Tracking AI: The capability is (not) near", presentation, 24th European Conference on Artificial Intelligence, Santiago de Compostela, Spain. [48]

Martínez-Plumed, F., E. Gómez and J. Hernández-Orallo (2020), "Tracking the impact and evolution of AI: The AIcollaboratory", Evaluating progress in AI, First International workshop, European Conference on Artificial Intelligence, Santiago de Compostela, Spain. [47]

Martínez-Plumed, F. et al. (2020), "Does AI qualify for the job: A Bidirectional model mapping labour and AI intensities", *Proceedings of the AAAI/ACM Conference on AI, Ethics and Society*, pp. 94-100. [22]

Morra, S. et al. (2012), *Cognitive Development: Neo-Piagetian Perspectives*, Psychology Press, Hove, UK. [16]

Mueller, S. (2010), "A partial implementation of the BICA cognitive decathlon using the Psychology Experiment Building Language (PEBL)", *International Journal of Machine Consciousness*, Vol. 2/2, pp. 273-288. [43]

Mueller, S. et al. (2007), "The BICA cognitive decathlon: A test suite for biologically-inspired cognitive agents", *Proceedings of behavior representation in modeling and simulation conference*. [44]

Piaget, J. (1964), "Cognitive development in children", *Journal of Research in Science Teaching*, Vol. 2/3, pp. 176-186. [15]

Piaget, J. (1936), *La naissance de l'intelligence chez l'enfant*, Delachaux et Niestlé, Lonay, Switzerland. [14]

Puentedura, R. (2006), "Transformation, technology and education", presentation, Strengthening Your District Through Technology workshop, 18 August, Maine School Superintendents Association, http://hippasus.com/resources/tte/. [55]

Purves, C., C. Cangea and P. Veličković (2019), "The PlayStation reinforcement learning environment (PSXLE)", *arXiv preprint arXiv*, Vol. 1912.06101. [56]

Rahwan, I. (2019), "Machine behaviour", *Nature*, Vol. 568/7753, pp. 477-486. [4]

San Antonio, T. (ed.) (1993), *Bayley Scales of Infant Development*. [24]

Schlangen, D. (2019), "Language tasks and language games: On methodology in current natural language processing research", *arXiv prepreint1398arXiviv*, Vol. 1908.10747. [31]

Schuelke, M. and E. Day (2012), "Ability determinants of complex skill acquisition", *Encyclopedia of the Sciences of Learning*, Springer. [12]

Seel, D. (2012), "Skill", *Encyclopedia of the Sciences of Learning*, Springer. [57]

Silver, D. et al. (2016), "Mastering the game of Go with deep neural networks and tree search", *Nature*, Vol. 529/7587, pp. 484-489. [9]

Taylor, A. (2018), "The Automation Charade", *Logic*, Vol. 5/1 August, https://logicmag.io/failure/the-automation-charade/. [3]

Tolan, S. et al. (2021), "Measuring the occupational impact of AI: tasks, cognitive abilities and AI benchmarks", *Journal of Artificial Intelligence Research*, Vol. 71, pp. 191-236. [23]

Vinyals, O. et al. (2017), "Starcraft II: A new challenge for reinforcement learning", *arXiv preprint arViv*, Vol. 1708.04782. [58]

Von Ahn, L. (2008), *Human computation*, presentation to IEEE 24th International Conference on Data Engineering, 7-12 April, Cancun. [2]

Wang, A. et al. (2019), "Superglue: A stickier benchmark for general-purpose language understanding systems", *arXiv preprint arXiv*, Vol. 1905.00537. [59]

Wasserman, E. and T. Zentall (2006), *Comparative Cognition: Experimental Explorations of Animal Intelligence*, Oxford University Press. [19]

Weber, B. (1997), “Computer defeats Kasparov, stunning the chess experts”, 5 May, New York Times. [7]

Zellers, R. et al. (2018), “Swag: A large-scale adversarial dataset for grounded commonsense inference”, *arXiv preprint arXiv*, Vol. 1808:05236. [42]

Zellers, R. et al. (2019), “Hellaswag: Can a machine really finish your sentence?”, *arXiv preprint arXiv*, Vol. 1905.078301400. [32]

## Notes

[1] The terms capability and capacity will be used more broadly, while the terms skill, knowledge and ability have more precise uses as follows. A skill is the "overlearned behavioural routine resulting from practice" (Seel, 2012[57]), which is represented by “the level of proficiency on specific tasks. It is the learned capability of an individual to achieve desired performance outcomes (Fleishman, 1972[52]). Thus, skills can be improved via practice and instruction” (Schuelke and Day, 2012[12]). Knowledge is typically used in a similar sense as skills but assumes a more theoretical or conceptual nature as opposed to the practical or actionable nature of skills. An ability "refers to a general trait, reflecting the relatively enduring capacity to learn tasks. Although fairly stable, ability may change over time primarily in childhood and adolescence through the contributions of genetic and developmental factors" (Schuelke and Day, 2012[12]). New skills are acquired using cognitive abilities and can build on previous skills and knowledge. See also section 3.

[2] The “Substitution, Augmentation, Modification, and Redefinition” (SAMR) model is a popular taxonomy covering the ways in which technology may affect tasks (Puentedura, 2006[55]; Hamilton, Rosenberg and Akcaoglu, 2016[53]), but the opposition of enhancement vs transformation does not work well for AI, especially as enhancement based on AI is usually coupled with significant modification and redefinition of tasks.

[3] A model or system overfits when it shows good performance for the examples seen during training (or examples from the training distribution) but generalises poorly for examples that are different from those seen during training.

[4] This has recently been rephrased as performance versus competence (Firestone, 2020[51]).

[5] www.onetcenter.org/taxonomy.html

[6] www.bls.gov/soc/2018/major_groups.htm

[7] www.ilo.org/public/english/bureau/stat/isco/

[8] https://ec.europa.eu/esco

[9] Summaries of these sources can be found in several tables and figures in Chapters 3-5 of (Hernández-Orallo, 2017[11]).

[10] This analysis excludes non-cognitive behavioural features, such as personality traits, from the analysis, as their extrapolation to AI systems is even more farfetched. This does not mean that personality in machines has not been studied or even tried to be measured. For more information about non-cognitive behavioural features in machines, see (Hernández-Orallo, 2017[11]).

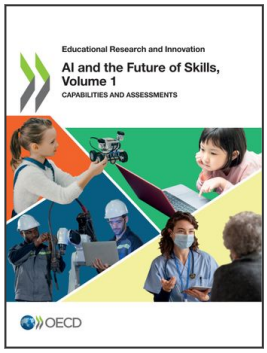
[11] For a summary of intelligence tests, developmental tests and attainment tests for humans, see Chapters 3 and 12 of Hernández-Orallo (2017[11]).

[12] Most of these tests are not freely available. This is partly for commercial reasons and partly because having the questions in advance would lead to specialisation: humans would prepare for the test.

[13] This has happened from CIFAR10 (image classification) to CIFAR100 (Krizhevsky, 2009[54]), SQuAD1.1 (Q&A) to SQuAD2.0, GLUE (language understanding) to SUPERGLUE (Wang et al., 2019[59]), Starcraft (real-time strategy) to Starcraft II (Vinyals et al., 2017[58]) and the Atari Learning Environment (ALE) (Machado et al., 2018[17]) to the PlayStation Reinforcement Learning Environment (PSXLE) (Purves, Cangea and Veličković, 2019[56]).

[14] https://mosaic.allenai.org/projects/mosaic-commonsense-benchmarks

[15] https://dynabench.org/




**From:**

# AI and the Future of Skills, Volume 1

## Capabilities and Assessments

**Access the complete publication at:**
https://doi.org/10.1787/5ee71f34-en

**Please cite this chapter as:**

Hernández-Orallo, José (2021), "Identifying artificial intelligence capabilities: What and how to test", in OECD, *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/85aeb432-en