15. Assessing artificial intelligence capabilities

Guillaume Avrin, artificial intelligence

As artificial intelligence (AI) becomes more mature, it is increasingly used in the world of work alongside human beings. This raises the question of the real value provided by AI, its limits and its complementarity with the skills of biological intelligence. Based on evaluations of AI systems by the Laboratoire national de métrologie et d'essais in France, this chapter proposes a high-level taxonomy of AI capabilities and generalises it to other AI tasks to draw a parallel with human capabilities. It also presents proven practices for evaluating AI systems, which could serve as a basis for developing a methodology for comparing AI and human intelligence. Finally, it recommends further actions to progress in identifying the strengths and weaknesses of AI vs. human intelligence. To that end, it considers the functions and mechanisms underlying capabilities, taking into account the specificities of non-convex AI behaviour in the definition of evaluation tools.

Introduction

Based on evaluations of AI systems by the Laboratoire national de métrologie et d'essais (LNE) in France, this chapter proposes a high-level taxonomy of AI capabilities. It then generalises this taxonomy to other AI tasks to draw a parallel with human capabilities. It also presents proven practices for evaluating AI systems, which could serve as a basis for developing a methodology for comparing AI and human intelligence. Finally, it recommends further actions to identify the strengths and weaknesses of AI vs. human intelligence. To that end, it considers the functions and mechanisms underlying capabilities, taking into account the specificities of non-convex AI behaviour in the definition of evaluation tools.

The chapter uses the terms "evaluation" and "evaluation campaign". An evaluation is a single test that aims to measure the characteristics (performance, explainability, etc.) of an intelligent system. Conversely, an evaluation campaign represents the process of evaluating products either vertically (by observing a range of products at a given time) and/or horizontally (by observing the evolution of the product over time).

Disciplinary field of artificial intelligence evaluation

This section provides a framework for LNE's "evaluation" activities in the AI field, while presenting the good practices acquired since this activity was set up in 2008.

The good practices at the heart of LNE evaluation campaigns are mainly the result of the search for a compromise between realism and reproducibility of experiments. It has led to the identification of features to be presented in campaigns below. Some of these features relate to the general organisation of the evaluation campaign, while others are more specialised on the evaluation process.

- Scientific

Evaluation campaigns preserve the demonstration aspect typically associated with them. However, they are based on the scientific criteria of assessment objectivity, performance measurement repeatability and experiment reproducibility. They also respect the requirements imposed by metrological rigour.

- Benchmark-based

The intelligent systems are evaluated through benchmarks. This means they perform well-specified tests in realistic environments or on databases. In addition, their performance is assessed by applying quantitative metrics.

- Modular

It is often not satisfactory to evaluate only the robot as a whole. Thus, the elements constituting the robot's architecture are broken down into functionalities (e.g. obstacle detection). These are then combined to perform more complex tasks (e.g. semantic navigation). The evaluation thus consists in Functionality Benchmarks (FBMs) and Task Benchmarks (TBMs). FBMs evaluate specific capabilities with a limited utility when used alone, while TBMs evaluate more complex activities (see below).

- Periodical

Evaluation campaigns should be organised as recurring events offering a similar evaluation framework each time (similar testbeds, similar testing datasets, same evaluation tools, etc.). This framework enables monitoring of the technological progress of the community of developers as a whole.

- Structured

Evaluation campaigns are structured to optimise effort and maximise impact. As such, they provide the scientific community with a stable set of benchmarking experiments. This, in turn, enables objective comparison of research results and can act as the seed for the definition of standards.

- Synergic

Evaluation campaigns should build on the well-established framework originally created by RoCKIn[1] and Quaero[2] projects and subsequently validated, perfected and extended by RockEU2[3], SciRoc[4], ROSE[5] and METRICS[6] projects.

- Open

Evaluation tools and annotated datasets should be publicly available. This will enable research and industry to develop and fine-tune their own algorithms, systems and products. Existing and prospective actors gain access both to difficult-to-obtain data with associated ground truth and to validated evaluation tools. Importantly, these by-products benefit the evaluator and promote the long-term sustainability of its evaluation campaigns. Users of the open data and tools will naturally be inclined to participate in the campaigns, thus creating a virtuous circle enabling their success.

Functionality benchmarks and task benchmarks

Evaluation campaigns include two groups of benchmarks (Amigoni et al., 2015[1]; Avrin, Barbosa and Delaborde, 2020[2]).

Functionality benchmarks (FBMs)

A functionality is conventionally identified as a self-contained unit of capability, which is too low level to be useful on its own to reach a goal (e.g. self-localisation, crucial to most applications but aimless on its own). A single component or a set of components can provide a functionality, and usually involves both hardware and software.

An FBM is a benchmark that investigates the performance of a robot component when executing a given functionality. A functionality is as independent as possible of the other functionalities of the system. In this way, functionality can be controlled as the sole dependent variable in the evaluation.

Task benchmarks (TBMs)

A task is an activity of a robot system that, when performed, accomplishes a goal considered useful on its own. A task always requires multiple functionalities to be performed. Finding and fetching an object, for example, involves functionalities such as self-localisation, mapping, navigation, obstacle avoidance, perception, object classification/identification and grasping. A TBM is a benchmark that investigates the performance of a robot system when executing a given task. TBMs are designed by focusing on the goal of the task, without constraining the means by which such goal is reached.

Evaluating the overall performance of a robot system while performing a task is interesting for assessing the global behaviour of the application. However, it does not allow evaluation of the contribution of each component. Nor does it put in evidence which components are limiting system performance.

On the other side, the good performance of each element in a set of components does not necessarily mean that a robot built with such components will perform well. System-level integration has, in fact, a deep influence on this, which component-level benchmarking does not investigate.

For these reasons, combining a TBM with FBMs focused on the key functionalities required by the task provides a deeper analysis of a robot system and better supports scientific and technical progress. The objective is to address the evaluation needs of end-users, integrators and equipment manufacturers.

Fairness of evaluation campaigns

This section looks at how to ensure an optimally fair treatment of the campaign’s participants. The notion of fairness is addressed in light of metrological considerations.

Simultaneity

The evaluator shall ensure a simultaneity of the evaluation, as required by the following considerations:

- **The difficulty to model the influence of environmental factors on the system's performance:** any outdoor experiment will never be completely repeatable. Clouds change in the sky; waves and tides modify visibility underwater, etc. This lack of repeatability, in addition to its influence on the metrological rigour of the evaluation, has an impact on the fairness of the evaluation between participating systems. It is not conceivable that one participant will have to operate in pouring rain, while another will suffer from maximum sunshine. In this regard, the evaluator shall define thresholds and limits in several parameters that are considered to influence performance of the devices. Outside of this acceptability range, the evaluator shall define remedial strategies to have intelligent systems compete in reasonably similar conditions.
- **The "*a priori* ignorance" imperative**: evaluated systems have a learning capability and consequently, should not have *a priori* knowledge of the testing environment (testbeds and testing datasets) used for the evaluation in order to avoid measurement bias and overfitting. This remark remains valid for systems that do not have learning skills since developers can influence the design of their systems if they have *a priori* information about testbeds and data.
- **The "*a posteriori* publication" imperative**: to ensure reproducibility of the evaluation experiments, testing environments used must be publicly described (and accessible if they are datasets) when the measurements and results are published.

This notion of "simultaneity" can sometimes be spread across the one or two days of the evaluation campaigns. The tolerance level about what may be considered "simultaneous" must, of course, be discussed on a case-by-case basis.

Impartiality

The evaluation must be carried out by a "trusted third party". This evaluator must have metrology expertise applied to the evaluated systems in order to develop an evaluation protocol common to all participants. In addition, it must guarantee there are no conflicts of interest between the campaigns' evaluator and participants.

Precise evaluation plan

Each evaluation must rely on an evaluation plan, a document that details the features of the following:

- one or more evaluation tasks that focus on a device or software performing a specification
- characteristics that need to be measured or estimated (performance, quality, safety, explainability, etc.)
- metrics (i.e. a formula that allows production of scores, such as accuracy, precision, recall, F-measure)
- test data or test environments (datasets or testbeds)
- evaluation tools (software for data collection, visualisation, comparison).

Evaluation task

The first step in organising an evaluation campaign is to specify and prioritise a set of evaluation tasks (FBMs and TBMs). They are deduced from the identification of scientific and technological barriers. The principal (campaign funder), who expresses the "business" need, defines the tasks rather than the evaluator (LNE and its potential partners). On the other hand, when a potential use case is identified, the evaluator must carry out the following checks:

- List solutions corresponding to the use case, with an estimate (when the information is accessible) of the performance limitations associated with their characteristics or conditions of use (costs, knowledge to be implemented for deployment, operation in highly constrained environments or for an extremely specific field, etc.).
- List the types of data required for the development and operation of such solutions, and their availability (considering regulatory or ethical limitations, the cost of collection, etc.).

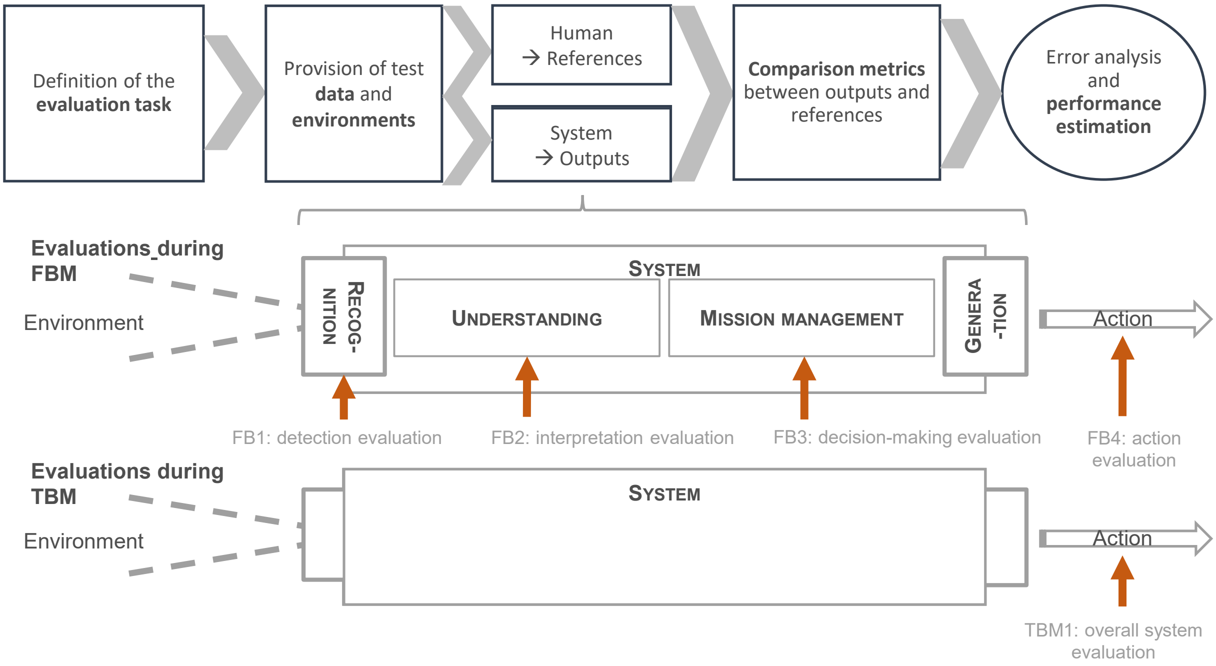
During this stage, the evaluator checks the feasibility of the campaign using the following criteria:

- possibility of objectifying the evaluation criteria
- difficulty of collecting and transmitting test data to the participants of the challenge (confidentiality of data inherent to use cases, availability of data, etc.), or making test environments available
- comparability of solutions for the use case (systems that can potentially take in extremely varied types of data may lead to significant adaptation of evaluation protocols, or even incomparability).

Evaluation method

The evaluation paradigm generally consists in comparing reference and hypothesis data. Reference data are the ground truth annotated by human experts or provided by measuring instruments in the test facility. Conversely, hypothesis data are the behaviour or output produced automatically by the intelligent system. This comparison allows the estimation of the performance, the reliability and other characteristics such as efficiency of robots. The evaluation can concern the entire system (during TBM) or the main technological components taken independently (during FBM), as shown in Figure 15.1.

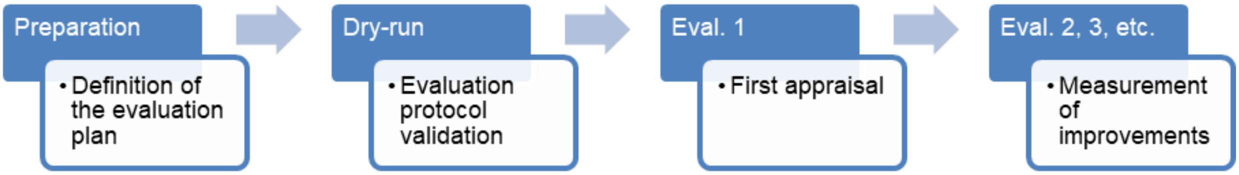
Figure 15.1. Evaluation method


Definition of the evaluation task
Provision of test data and environments
Human → References
System → Outputs
Comparison metrics between outputs and references
Error analysis and performance estimation
Evaluations during FBM
Environment
SYSTEM
RECOG-NITION
UNDERSTANDING
MISSION MANAGEMENT
GENERA-TION
Action
FB1: detection evaluation
FB2: interpretation evaluation
FB3: decision-making evaluation
FB4: action evaluation
Evaluations during TBM
Environment
SYSTEM
Action
TBM1: overall system evaluation

Some evaluation campaigns last several years and include several evaluations. The repeated evaluations allow the principal to assess the effectiveness of the funding granted for the organisation of the evaluation campaign. For example, this could estimate the performance of potential technological solutions that address its use case. For developers, repeated evaluations allow them to update the technological components of the intelligent system according to the quantitative results obtained.

The dry-run evaluation guarantees the smooth implementation of the campaign. It allows the evaluator to ensure its evaluation plan is both realistic with respect to the capabilities of the systems, and fair among the different technologies used by participants. Thus, the dry run can experiment with several test environments and metrics to define the best evaluation protocol that will be fixed during the official evaluation campaigns. Several official evaluation campaigns follow the dry run. These aim at objectively measuring the progress of participating robots in real field conditions. To this end, the evaluation plan is meant to be adapted throughout the campaign to accompany the evolutions of the participants' technological solutions. The steps of an evaluation campaign are presented in Figure 15.2.

Figure 15.2. Steps of a campaign involving several evaluations


Preparation
• Definition of the evaluation plan
Dry-run
• Evaluation protocol validation
Eval. 1
• First appraisal
Eval. 2, 3, etc.
• Measurement of improvements

Comparison metrics

The capability measurements must be quantitative and provided by a formula ("metric") that indicates the distance between the reference and the hypothesis, or measures capacity directly. The distance between the reference and the hypothesis could be measured by the distance between a real or ideal trajectory in a navigation task, the number of false positives and false negatives in an image recognition task, the binary success of a task, etc. A direct measurement could be time to completion, distance covered, etc.

Test data or test environments (datasets or testbeds)

Evaluations can be based on physical (in real or laboratory conditions on testbeds) and/or virtual testing environments (simulators and testing datasets). Pros and cons of the two types of environments are presented below.

Table 15.1. Strengths and weaknesses of test environments

	Physical testbeds	Simulators	Datasets
Exhaustiveness of tested scenarios	*	***	**
Realism of tested scenarios	***	*	**
Generation of new data	***	**	*
Dynamic and closed-loop testing	***	**	*
Experimentations reproducibility and measure repeatability	*	***	***
Cost of each test	*	***	**
Data obsolescence	***	**	*

Note: The number of asterisks indicates the rank of the test environment type: *** corresponds to the best solution and * to the worst.

Taxonomy of evaluations carried out by Laboratoire national de métrologie et d’essais

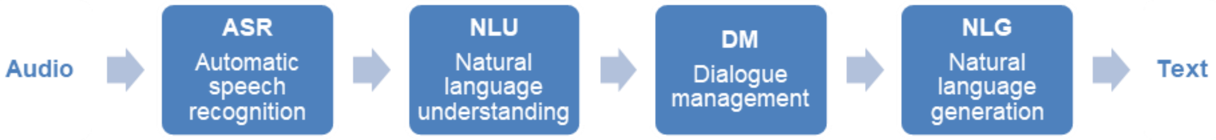
Clustering Laboratoire national de métrologie et d’essais’s artificial intelligence evaluations

LNE has carried out more than 950 evaluations of AI systems since 2008. These include areas such as language processing (translation, transcription, speaker recognition, etc.), image processing (person recognition, object recognition, etc.) and robotics (autonomous vehicles, service robots, agricultural robots, intelligent medical devices, industrial robots, etc.). Examples of evaluations in the context of research and development projects are presented in Table 15.2.

The evaluation tasks have been grouped into the capabilities of recognition, understanding, mission management and generation. These are an extension of the “sense-think-act” paradigm and an adaptation of the AI cycle “Perception – Learning – Knowledge representation – Reasoning – Planning – Execution” (Beetz et al., 2007).

These capabilities are also consistent with the NIST 4D/RCS reference architecture for autonomous vehicles (Albus, 2002[3]); with the “SPACE” breakdown into functions (Sense, Perceive, Attend, Apprehend, Comprehend, Effect action) that underpin intelligent behaviour (Hughes and Hughes, 2019[4]) and finally, with most cognitive architectures (Kotseruba and Tsotsos, 2016[5]; Ye, Wang and Wang, 2018[6]). These can be illustrated by the dialogue systems, for which such a division is common (see Figure 15.3) (Leuski and Traum, 2008[7]).

Figure 15.3. Division of dialogue systems capabilities


Audio
ASR
Automatic speech recognition
NLU
Natural language understanding
DM
Dialogue management
NLG
Natural language generation
Text

Comparative evaluation of different architectures

LNE also assesses another capability of AI, which cuts across the different capabilities listed above: the system's capability to learn and update its parameters throughout its lifecycle.

The differences between the cognitive architectures cited above do not result so much from divergent points of view on potential capabilities. Rather, they reflect two other factors. First, they were developed in different contexts (different perimeters and objectives of the associated research projects). Second, they have different hypotheses regarding the neural processes underlying these functionalities (symbolic, connectionist or hybrid architectures, centralised or decentralised processing, etc.). Summaries of the main cognitive architectures[7] and the main capabilities[8] covered by these architectures are available in the literature.

The comparative evaluation of these different architectures is a vast subject of research. The evaluation criteria considered include the generality of the architecture. This measures the types of tasks and environments that can be handled by systems developed according to this architecture. This measurement, in turn, is assessed in terms of versatility and taskability.

Versatility is defined as the number of ways in which the system designed according to the architecture can solve the same task, using different capabilities. Meanwhile, taskability is the number of different tasks that can be performed by the system receiving external commands. This generality feature of an architecture is directly related to the general intelligence of the resulting systems (Langley, Laird and Rogers, 2009[8]) and therefore directly relevant to this study.

There is a wide variety of terms within these cognitive architectures to describe their capabilities. The first column of Table 15.2 proposes a first equivalence between these terms.

These cognitive architectures are consistent with each other. They are not only interested in reproducing the external behaviour of biological intelligences but also in modelling the internal properties of their cognitive systems. They propose a blueprint for cognitive agents depicting the arrangement of functional units. This facilitates implementation of their principles in mechatronic systems [see "*architecture-as-methodology*" in Jiménez et al. (2021[9])].

These cognitive architectures also provide a formalism for presenting human capabilities (and dealing with the intrinsic complexity of cognitive systems) that can be reproduced in artificial systems.[9] In this way, they represent a bio-inspired and integrated taxonomy of human and artificial capabilities and they facilitate the comparison of these capabilities when they are implemented in humans and in machines.

Exclusivity and exhaustiveness

As these cognitive architectures are used to assemble technological components of intelligent mechatronic systems, each capability can be associated to an exclusive component or group of components. Mutual exclusivity between these capabilities is thus guaranteed. For obvious cost reasons, engineers using a cognitive architecture to design their intelligent systems would have no interest in building in functional redundancy between the different components (which must be distinguished from the redundancy intended to meet the safety requirements of critical systems).

This modular architecture therefore makes it possible to isolate the technological components that underpin the different capabilities (i.e. the functional units) and to carry out input-output evaluations on each component to evaluate each capability independently.

Various studies have investigated the exhaustiveness of the capabilities covered by these architectures. However, there does not yet seem to be a consensus regarding the most comprehensive architectures. Some prefer CLARION and AIS (Kotseruba and Tsotsos, 2016[5]), while others prefer OpenCogPrime (Ye, Wang and Wang, 2018[6]). Exhaustiveness can be measured by the "generality", i.e. the number of tasks and environments in which a system built according to this architecture can be used.

Table 15.2. LNE's evaluation tasks and metrics

Automatic information processing systems			
Task	**Capability**	**Metrics**	**Project**
Speaker verification for criminalistics application	Recognition	Equal error rate (EER), Detection cost function (DCF), Detection error trade-off (DET), Probabilistic linear discriminant analysis, etc. (Ajili et al., 2016[10])	FABIOLE (2013-16) VOXCRIM (2017-21)
Automatic speech recognition	Recognition	Word error rate (WER), Automatic transcription evaluation for named entities (ATENE), Word Information Loss (WIL), Relative information loss (RIL), IN, Near (Ben Jannet et al., 2015[11])	VERA (2013-15)
Speaker diarisation	Recognition	Diarisation Error Rate (DER) (Prokopalo et al., 2020[12])	ALLIES (2017-20)
Speaker diarisation across time	Recognition Learning	Average DER across audio file, weighted by duration of the file	
Lifelong learning diarisation	Recognition Learning	The DER is computed on the final version of the hypothesis for each document penalised by the cost of interacting with the user in the loop	
Translation, translation across time, lifelong translation	Recognition Understanding Generation Learning	Bilingual evaluation understudy (BLEU) adaptations similar to those of the speaker diarisation tasks	
Recognition of patients' vital signs (breathing, heart rate, etc.).	Recognition	Estimated global error rate (EGER), Precision, Recall, F-measure	AIR (2020-22)
Transcription from TV feeds	Recognition Generation	Word error rate (WER), (Galibert et al., 2014[13])	ETAPE (2010-12)
Named entity recognition (detection, classification, decomposition)	Recognition	Entity Tree Error Rate (ETER), Slot error rate (SER), Error per response (ERR), EDT value, Local entity detection and recognition (LEDR) (Ben Jannet et al., 2014[14])	QUAERO (2008-14)
Question-answering systems	Recognition Understanding Mission management Generation	QA distance measure, (Bernard et al., 2010[15])	
People recognition in multimodal conditions	Recognition	EGER (Kahn et al., 2012)	REPERE (2012-14)
Translation of newspapers articles and broadcast news transcriptions that come from various radio and television programmes	Recognition Understanding Generation	Translation Error Rate (TER), BLEU, Human-mediated translation edit rate (HTER)	TRAD (2012-14)
Area segmentation	Recognition	ZoneMap, Pset, DetEval, Jaccard, (Brunessaux et al., 2014[16])	MAURDOR (2012-14)
Identification of the writing type (handwritten, printed, unspecified)	Recognition	Accuracy	
Language identification	Recognition	Accuracy	
Text-to-text transcription and optical character	Recognition	WER, Character error rate	

recognition		(CER)	
Extraction of logical structure (logical connections between semantic areas)	Recognition Understanding	Precision, Recall, F-measure (Oparin, Kahn and Galibert, 2014[17])	
Synthesis of video and text information	Recognition Understanding Mission management Generation	WER, SER, Precision, Recall, F-measure	IMM (2013-16)
Automatic speech recognition performance prediction	Recognition Understanding	Mean Absolute Error (MAE) and Kendall (Elloumi et al., 2018[18])	Autre – 2018
Satellite image classification	Recognition	EGER, ZoneMap, Jaccard	Confidential (2019-20)
Recognition of aircraft movement patterns from radar data	Recognition	EGER, Precision, Recall, F-measure	Confidential (2019-20)
Transcription by smartphone intelligent personal assistant	Recognition	WER, Accuracy	Confidential (2019)
QA by smartphone intelligent personal assistant	Recognition Understanding Mission management Generation	Accuracy	
Robotic systems			
Task	**Capability**	**Metrics**	**Project**
Crops and weeds recognition	Recognition	EGER, Precision, Rappel, F-measure (Avrin et al., 2019[19]); (Avrin et al., 2020[20])	Challenge ROSE (2018-21)
Mechanical/electrical weeding action	Generation	Accuracy	
Full agricultural weeding robot evaluation	Recognition Understanding Mission management Generation	Accuracy	
Advanced driver assistance (ADAS)	Recognition Understanding Mission management Generation	Time to collision, time exposed to time to collision, time to brake, time to steer, time to react	SVA/3SA (2015-22)
Climbing up 10 cm high stairs without handrail, climbing up 15 cm high stairs with handrail, walking over stepping stones, walking on a beam, walking on a flat ground, walking on a slope, walking over obstacles	Recognition Understanding Mission management Generation	Walked distance, success rate, max tracking error, duration of the experiment, etc. (Stasse et al., 2018[21])	Robocom++ (2017-20)
Human detection for logistics robots	Recognition	EGER, Precision, Recall, F-measure	Blaxtair Safe (2019-20)
Estimate the stopping distance (conventional or emergency) under load and maximum speed	Generation	Linear distance measurement	ECAI (2019-20)

As shown in Table 15.2 the same task may involve one or more capabilities depending on the context. For example, an information retrieval task may rely only on the mission manager if the information is stored in memory. It may require recognition and understanding if it involves searching for information in text. A medical diagnosis may be based solely on a capability for recognition, or may also involve a phase of reasoning. A medical prescription will involve the "mission manager" component.

Generalisation to other artificial intelligence tasks

The capabilities presented in the previous section are defined in more detail in Table 15.3 and generalised to other typical AI tasks. Table 15.4 illustrates the presence of these capabilities in AI systems. Table 15.5 provides an example of how to implement the evaluation process to assess these capabilities for a specific AI system.

Table 15.3. AI capabilities generalisation

AI capabilities (and equivalent words)	Examples of AI tasks	Example of AI output
Recognition: perception/acquisition of sensory information (vision, hearing, etc.)	Speech recognition, optical character recognition, tokenisation, named entity recognition, lemmatisation, parsing, pose estimation, face verification, scene segmentation, person reidentification, image classification, etc.	"object: glass", "position: falling" and "object: human arm", "position: stretched"
Understanding: contextualisation, interpretation, comprehension, conceptualisation, assimilation (relating to system state, storage, etc.)	Knowledge representation, 2D/3D mapping, information extraction, image captioning, etc.	"the human tries to catch the falling glass"
Mission manager: decision making, cogitation, cerebration, reasoning, inferring, arbitration (judgement), etc.	Prediction, planning, optimisation, selection between different options, self-check, etc.	Identification of the best trajectory to catch the glass safely before it touches the ground
Generation: action	Navigation, speech synthesis, locomotion, manipulation (grasping, etc.), content generation (image, etc.), etc.	Generation of the movement of the robotic arm and the gripping effector to catch the glass
Learning: adaptation, knowledge storing	Parameters update (supervised, unsupervised, reinforcement learning, etc.), operation algorithm change.	If "broken glass": failure, update the trajectory generation parameters, otherwise do nothing.

Table 15.4. Examples of AI capabilities for different tasks

Recognition	Understanding	Mission management	Generation
Autonomous car			
Traffic-sign recognition, obstacle recognition, etc.	Velocity synthesis, image plan mapping, relationship identification, etc.	Motion planning, risk assessment, etc.	Vehicle control, braking, steering, driver alert, etc.
Text summarisation			
Sentence segmentation, word segmentation, feature extraction	Feature frequency, similarity computation, sentences comparison and scoring	Sentences selection and assembly	Summary generation
Recommendation systems			
Analysis of rating	Analysis of behaviour, contextualisation based on location, time, user profile, etc.	Comparison to other user preferences	Recommendation of objects.

Table 15.5. Example of the evaluation steps for autonomous weeding robots (from ROSE and METRICS projects)

Step	Detail
Formalisation of the need	• What is the objective: autonomous weeding of the intra-row of agricultural plots. • Which crops should be considered in priority: corn and beans. • What are the weeds to be considered in priority: lamb's quarter, matricaria, ryegrass and wild mustard.
Feasibility analysis	• Mapping of weed control robots on the market. • Identification of the main capabilities useful for the task (weed detection, weeding decision making, weeding action). • Estimation of the costs associated with the evaluation of these different capabilities: cheap weed images to produce, expensive test farm to set up, etc.
Formalisation of the evaluation tasks	• Recognition: segmentation of weeds and crops on images. • Generation: navigation, weeding action. • Etc.
Formalisation of the evaluation criteria and metrics	• Segmentation metrics: estimated global error rate (EGER), Jaccard index, Zonemap, etc. • Generation metrics: biomass estimation, counting of weeds removed. • Etc.

Relevance of the proposed capability taxonomy

Relevance to artificial intelligence

This section reviews the mutually exclusive and collectively exhaustive capabilities (MECE character) of the different taxonomies to assess their relevance.

Exclusivity

The taxonomy proposed in the previous section is inspired (although simplified) by cognitive architectures. These are designed to assemble different functional units (each representing its own capability) to form an information processing pipeline. As each unit has its own function, these cognitive architectures are designed to ensure the mutually exclusive nature of the capabilities. In this way, they avoid any redundancy that would be detrimental in terms of the manufacturing cost of the AI system. However, with the rise of end-to-end learning (Shibata, 2017[22]), the boundary between these different functions is blurring as design moves from this traditional "pipeline".

Exhaustiveness

The proposed taxonomy seems to cover the capabilities of the main cognitive architectures, although with a high level of abstraction (Kotseruba and Tsotsos, 2016[5]; Ye, Wang and Wang, 2018[6]; Hughes and Hughes, 2019[4]). High-level capabilities could be further broken down into tasks, while retaining their MECE nature. Table 15.6 provides an example of the decomposition of a high-level capability, which is modality- and application-independent, into modality-dependent tasks and application-dependent sub-tasks. This division can be continued until specific tasks are reached (such as the manufacturing tasks proposed in Huckaby and Christensen (2012[23]): place, transport, retract, slide, insert, pick up, align, etc.).

Table 15.6. Example of breaking down the recognition capability into sub-tasks

Capability (modality- and application-independent)	Modality-dependant task	Modality- and application-dependant sub-task
Recognition	Image recognition	Optical character recognition
		Face recognition
		Pose estimation
		Etc.
	Language recognition	Tokenisation
		Lemmatisation
		Named entity recognition
		Etc.
	Etc.	Etc.

The breakdown of capabilities proposed for the taxonomy is also relevant given that substantial progress on a task in one capability advances AI performance on other associated tasks (see Table 15.2 for examples of tasks for each capability). This is, in particular, the consequence of the democratisation of the use of pre-trained algorithms and inductive transfer (Moon, Kim and Wang, 2014[24]).

This observation is even more striking for a particular modality related to a given capability [e.g. visual recognition (Razavian et al., 2014[25])or speech recognition (Howard and Ruder, 2018[26]; Peters et al., 2018[27]; Devlin et al., 2019[28])].

This is the case in part because the tasks for a given capability usually involve the same types of algorithms. For example, recognition tasks typically use classification, clustering or mapping algorithms. Conversely, mission management tasks will use more optimisation or regression algorithms. These correspondences between types of tasks to be automated and types of algorithms used for automation are discussed further below.

These dependencies between progress on tasks associated with the same capability are much more evident between high-level tasks and their sub-tasks. In particular, some work highlights the critical implications that progress in certain sub-tasks can have for AI as a whole (Cremer and Whittlestone, 2020[29]).

Relevance to humans

If the proposed taxonomy seems relevant to AI, another question arises: will it allow an effective comparison between human and AI capabilities? The answer requires two considerations.

First, this taxonomy is related to cognitive architectures. As such, they already provide an integrated view of human and artificial capabilities, with particular caution regarding jingle-jangle fallacies mentioned in Primi et al. (2016[30]). Indeed, such a taxonomy should be independent of the underlying methods and equipment used Shneier et al. (2015[31]).

Second, the idea of decomposing high-level capabilities into a pipeline of lower-level capabilities also seems relevant for the analysis of human capabilities. Tolan et al. (2020[32]) highlight this type of dependence between high-level capabilities and lower-level skills. This pipeline decomposition is also consistent with the levels of autonomy proposed in Huang et al. (2007[33]) to characterise the assistance of the machine to the human and vice versa.

The decomposition choices, of which a first example is provided in Table 15.6, are in turn complex to perform. A consensus seems to be found in the idea of starting the taxonomy with high-level capabilities that are non-specialised (Hernández-Orallo, 2017[34]). Neubert et al. (2015[35]) called these capabilities with a higher level of abstraction “Core domain skills”, “Transversal skills” and “Basic cognitive skills”, while O*NET[10] refers to them as “cross-occupational activities”. Chapter 7 explores these skills in more detail.

The question of correspondence with the taxonomies of human capabilities also arises (Hernández-Orallo, 2017[34]; Hernández-Orallo, 2017[36]; Tolan et al., 2020[32]). An association is proposed in Table 15.7.

Table 15.7. Correspondence between AI and human capabilities

Human abilities	Capabilities
Memory processes	Mission management Learning
Sensorimotor interaction	Recognition Understanding Mission management Generation
Visual processing	Recognition
Auditory processing	Recognition
Attention and search	Recognition
Planning and sequential decision making and acting	Mission management Generation
Comprehension and compositional expression	Understanding Mission management Generation
Communication	Mission management Generation
Emotion and self-control	Mission management Generation
Navigation	Mission management
Conceptualisation, learning and abstraction	Understanding Learning
Quantitative and logical reasoning	Mission manager
Mind modelling and social interaction	Understanding Mission manager Generation
Metacognition and confidence assessment	Mission manager

Source: Hernández-Orallo, (2017[36]).

The human capabilities shown in Table 15.7 are mainly inspired by psychometrics, comparative psychology and cognitive science. They correspond to combinations of different capabilities proposed for AI, although the proposed taxonomy has a high level of abstraction. As a consequence, transcribing these human capabilities into an AI system would require different functional units. These capabilities would be called "composite". In AI, composite capabilities are complex to evaluate. The modular organisation of capabilities within cognitive architectures instead allows each technological component to be evaluated independently, through input-output evaluations, as discussed in Section 3.

Relevance of evaluation methods to compare human and artificial capabilities

Relevance of artificial intelligence tests

This chapter presents an approach used by LNE to evaluate AI systems based on the implementation of benchmarks (i.e. standard tests). The test-based approach is also commonly used to assess human capabilities. School exams and neuropsychological evaluations (perceptual, motor, attentional tasks, etc.) rely on tests. Moreover, the *a priori ignorance, a posteriori publication and impartiality* requirements are equally important for such human dedicated tests. Even the adaptive/adversarial testing approaches used for AI have their equivalent for human testing. Adaptive testing is found in GRE, as well as in oral tests such as the one used by German dual vocational education and training (see Chapter 9).

Since the test-based approach is already used to evaluate both biological and artificial capabilities, it would be interesting to compare these competences. In most of the LNE data-based evaluations, humans perform the reference annotations against which the outputs of the intelligent system under evaluation are compared (see sub-section "Precise evaluation plan"). In practice, several humans annotate each piece of data in the test database[11] to carry out inter- and intra-annotations agreement analyses (Mathet et al., 2012[37]) and to verify the ground truth associated with the test data. Therefore, most evaluations of AI systems include, from the beginning, a comparison with humans.

Tests designed for AI are also interesting because they are modular (cf. sub-section "Disciplinary field of AI evaluation"). As well, the evaluation tasks (task benchmarks and functionality benchmarks) follow the division of human capabilities into functional units proposed by cognitive architectures (cf. sub-section "Clustering LNE's AI evaluations"). Thus, they are optimal to compare human and artificial capabilities.

For these reasons, tests specifically designed for AI systems could occupy a prominent place in the OECD's *Artificial Intelligence and the Future of Skills* project.

Relevance of human tests

Many tests designed for humans seem unsuitable for AI.

First, tests are generally conducted with environments whose size (questionnaire, duration of driving licence exams, etc.) is not adapted to the specifics of AI behaviour. Indeed, AI behaviour is largely non-convex and non-linear. It is not possible to evaluate its performance at a few points and deduce by interpolation and extrapolation its performance on the whole operating domain. Thus, testing environments are set up to maximise the exhaustiveness of the test scenarios covered (e.g. virtual testing). On the contrary, humans have much less chaotic behaviour. This is why a driving exam of less than 60 minutes, or a written test with about 20 questions, is sufficient to test a human's performance.

Second, they sometimes focus on tasks (e.g. IQ tests) that can be easily overfitted by AI. Conversely, the risk of human overfitting of tasks designed to evaluate AIs seems much lower.

Third, LNE has never evaluated some human capabilities presented in Table 15.7 in AI. Perhaps the task was not immediately relevant to the machine kingdom (e.g. it has no "self-control"). Or perhaps it was not evaluated as part of a specifically dedicated task, even it was a sub-component of a more complex task being evaluated (e.g. memory processes, quantitative and logical reasoning). As another possibility, no client may have ever asked LNE to assess this capability (e.g. "Emotion", "Mind modelling and social interaction").

This third finding is informative for the OECD study because it may indicate one of two things:

- AI is too immature to perform this task. Therefore, there is no system on the market that can perform it and useless to organise an evaluation campaign for it.
- Economic stakeholders have not yet deemed the assessment of this capacity as useful.

The latter does not necessarily mean the automation of this capability has no market value. Indeed, most often only the "critical" systems incorporating AI (which present a risk to goods and/or people) are assessed by trusted third parties such as the LNE, in line with European regulations.[12]

Finally, human tests are designed to assess abilities, some of which have a name that may be questionable for AI. A somewhat simplistic understanding of the "memory" capability in Table 15.7, for example, could suggest this task is not relevant for AI, since AI never forgets. On the contrary, if this task concerns the ability to store, recognise and re-use knowledge in general, then it seems a critical step not yet reached in AI development (Cremer and Whittlestone, 2020[29]).

Similarly, many tasks automated by AI, such as optimising movements on a farm plot to weed a maximum of weeds in a minimum of time call for "quantitative and logical reasoning" skills (Avrin et al., 2020[20]; Avrin et al., 2019[19]). However, it is not clear whether this task is more consistent with this capability than with "planning and sequential decision making and acting" or even "navigation".

Relevance of test methods specific to the intelligences being tested

With respect to the non-convexity of AI behaviour and the convexity of human behaviour, and given the risks of overfitting, evaluation tools must generally be defined according to the intelligence to be evaluated. Two elements generally define the testing tools (measuring instrument, test dataset, etc.) to be used in an evaluation. First, there are the expected functionalities (image recognition, scene understanding, etc.) of the evaluated intelligent system. Second, there are the technological solutions underpinning these functionalities, be they algorithms (CNN, SVM, etc.) or biological neural networks.

Another taxonomy relating to the type of technical solution (algorithms, biological neural architectures, etc.) used to achieve the functionality could therefore be established. This "mechanisms taxonomy" would be used to define the test protocol used (sampling and number of tests/questions, etc.) to evaluate the skills listed in the "capabilities taxonomy" and offered by the intelligent system under study.

This does not mean that some systematic correspondences between the "capabilities taxonomy" and the "mechanisms taxonomy" cannot be found. For example, recognition tasks are often automated by deep learning algorithms. In addition, comprehension tasks often rely on knowledge graphs and mission management tasks on reasoners.

This conclusion, moreover, is quite logical with regard to certain specificities of AI and human intelligence:

- Other elements than capabilities can influence human performance, such as traits, interests and values (De Fruyt, Wille and John, 2015[38]). The socio-emotional characteristics of human performance must be considered when designing the test. This is not the case for AI.
- AI can be duplicated and simulations run in parallel to test a large number of test scenarios; it is not possible to do the same for humans.

Relevance of task assessments

Although the assessment of AI and human intelligence capabilities are the focus of the study, task-based assessments may still be useful given the two points below:

- **There is no single combination of capabilities to perform a given task**. Each type of agent will try to rely on its best capabilities: AI systems will rely on their remembering and retrieving skills, their unbounded working memory, their speed of calculation, their perfect attention span; humans will rely on their unrivalled manipulation skills, common sense reasoning, frugal learning skills, etc.
- **The end-to-end learning approach of AI can render obsolete/impossible the evaluation of certain capabilities** (e.g. it is not possible to evaluate the performance of an end-to-end dialogue system in named entity recognition).

Relevant commonalities between all test methods

The test-based evaluation approach is common to both AI and human intelligence. It seems to be a crucial avenue to compare them. The Animal-AI testbed is, for example, dedicated to the evaluation of non-specific capabilities in both animals and AIs. How could standard test modules, such as ASTM E2919-14 for "Pose measurement", be designed for AI in many different applications in manufacturing, construction, medicine and aerospace, to evaluate human performance?

In addition, the test-based approach has other attributes that can inspire the expert judgement-based method of this study:

- The assessments should be **modular** (in agreement with the taxonomical approach of the OECD project), as already discussed above.
- The **impartiality** of evaluations should be ensured: an expert could underestimate or overestimate the capabilities of AI systems due to a conflict of interest.

Recommendations

This chapter capitalises on LNE's experience in evaluating AI systems to address two main questions:

- Which taxonomy should be used to compare AI and human intelligence capabilities?
- What evaluation tools and methods should be used to compare these capabilities?

It proposed a first taxonomy, simple but relevant to both biological and artificial intelligences. It then made recommendations regarding assessments to compare these intelligences. To make progress in answering the two questions above, and to pursue the impulse launched by the OECD in a particularly constructive, methodical, concerted and transparent spirit, the following actions would be useful:

- **Classify human and AI capabilities in terms of functions and mechanisms**

Intelligent systems (human or machine) perform very different functions (e.g. face recognition and bipedal walking, medical diagnosis and navigation of an unmanned aerial vehicle) using information processing mechanisms that rely on the same elementary principles. Conversely, within the same category of functions, different mechanisms can be used (rational or intuitive channels for humans, neural networks or expert systems for AI). For AI, grouping by evaluation metrics, types of automated tasks (classification, segmentation, etc.) and types of algorithms used (CNN, SVM, etc.) are examples of interesting avenues.

- **Organise evaluation tools around this double classification (function and mechanism)**

The general architecture and the hardware devices of the test benches to be set up (input/output channels, feedback, real time, etc.) are closely related to the mission of the system to be evaluated. Conversely, protocols to be followed (sampling and annotation of the operating domain, number of tests, etc.) will be determined mainly by the cognitive or computer mechanisms involved. In a maths competition, for example, a grading scale and a reader are mobilised; in a singing or figure skating competition, a jury is set up; in a sitting trial, both a professional legal judge and a popular jury are involved.

- **Formalise the influence of the non-convexity and intra-task variability of behaviour on the evaluation tools to be implemented**

AI generally has a non-convex behaviour with significant intra-task performance variability, while humans have a convex and stable behaviour. The behaviour convexity has a direct impact on the evaluation methods. It constitutes a gap between AI and human testing approaches that makes any assimilation difficult at this stage, in either direction. The evaluator of an intelligent machine has no choice but to go through the operating domain in all its corners. It must be tested at each of its operating points with a sampling step that is immediately related to the extremely unstable, non-linear character of its reactions.

The evaluator of a human being will be much less precise. The evaluator will be satisfied with probing the acquisition of a know-how by putting the person in “typical” situations that solicit the various components of the competence (e.g. the driving licence exam vs. the long test campaigns of the autonomous vehicle). The evaluator thus hypothesises that the person has regulation capabilities and mental resources more general and common to the ordinary human being that will make him/her able to face any intermediate situation.

The machine does not have them yet. This is probably because of its specialisation and its relative simplicity. However, it is also undoubtedly because of the technologies and processes used, which are not, or not sufficiently, superimposable on the natural cognitive mechanisms, composite and articulated, inherited from evolution.

These major differentials – the instability of intelligent systems – are of course to be nuanced precisely according to these technologies and applications. This is the main criterion on which to base improvements of the proposed taxonomy for comparison and cross-fertilisation between the two disciplines.

- **Deepen the discussion concerning the inter-task and intra-capability repercussions of the progress made in AI, to identify the root of AI capabilities and, by analogy, that of the human being**
- **Develop a broadly shared set of resources, methodologies and evaluation metrics that will enable these analyses to be conducted and AI/human progress to be tracked**

The strengths and weaknesses of human intelligence compared to AI by a technical and comparative rapprochement in terms of taxonomic and methodological unity of appreciation should be identified as soon as possible. This should accompany the progress in AI and cognitive sciences and, in particular, pilot what contributes to identify their "greatest common divisors".

AI seems to be the source of changes that are extremely favourable to the destiny of humanity, such as a radical emancipation from work. Therefore, this evolution should be supported by seeking to control the risks rather than pushing it back or slowing it down. Otherwise, humans will end up enduring AI without having prepared for it sufficiently.

References

Ajili, M. et al. (2016), "FABIOLE, a speech database for forensic speaker comparison", in *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 16*, European Language Resources Association, Luxembourg, https://www.aclweb.org/anthology/L16-1115.pdf. [10]

Albus, J. (2002), "4D/RCS: a reference model architecture for intelligent unmanned ground vehicles.", in *Proceedings Volume 4715, Unmanned Ground Vehicle Technology IV*, Aerosense 2002, Orlando, FL, https://doi.org/10.1117/12.474462. [3]

Amigoni, F. et al. (2015), "Competitions for benchmarking: Task and functionality scoring complete performance assessment", *IEEE Robotics & Automation Magazine*, Vol. 22/3, pp. 53-61, https://doi.org/10.1109/MRA.2015.2448871. [1]

Avrin, G., V. Barbosa and A. Delaborde (2020), "AI evaluation campaigns during robotics competitions: The METRICS paradigm", presented at the First International Workshop on Evaluating Progress in Artificial Intelligence - EPAI 2020, Santiago de Compostela, Spain, https://dmip.webs.upv.es/EPAI2020/papers/EPAI_2020_paper_5.pdf. [2]

Avrin, G. et al. (2020), "Design and validation of testing facilities for weeding robots as part of ROSE Challenge", presented at the First International Workshop Evaluating Progress in Artificial Intelligence of the European Conference on Artificial Intelligence, Santiago de Compostela, Spain, https://dmip.webs.upv.es/EPAI2020/papers/EPAI_2020_paper_6.pdf. [20]

Avrin, G. et al. (2019), "Boosting agricultural scientific research and innovation through challenges: The ROSE Challenge example", 3rd RDV Techniques AXEMA, SIMA, https://www.seaperch.org/challenge. [19]

Ben Jannet, M. et al. (2014), "ETER: A new metric for the evaluation of hierarchical named entity recognition", in *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, European Language Resources Association, Reykjavik, https://www.aclweb.org/anthology/volumes/L14-1/. [14]

Ben Jannet, M. et al. (2015), "How to evaluate ASR output for named entity recognition?", in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2015-January*, International Speech Communication Association, Baixas, France. [11]

Bernard, G. et al. (2010), "A question-answer distance measure to investigate QA system progress", in *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, European Language Resources Association, Malta, https://www.aclweb.org/anthology/volumes/L10-1/. [15]

Brunessaux, S. et al. (2014), "The Maurdor Project: Improving automatic processing of digital documents", *11th IAPR International Workshop on Document Analysis Systems*, pp. 394-354, http://dx.doi.org/10.1109/DAS.2014.58. [16]

Cremer, C. and J. Whittlestone (2020), "Canaries in technology mines: Warning signs of transformative progress in AI", *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 6, pp. 100-109, http://doi.org/10.9781/ijimai.2021.02.011. [29]

De Fruyt, F., B. Wille and O. John (2015), "Employability in the 21st century: Complex (interactive) problem solving and other essential skills", *Industrial and Organizational Psychology-Perspectives on Science and Practice*, Vol. 8/2, pp. 276-U189, http://dx.doi.org/10.1017/iop.2015.33. [38]

Devlin, J. et al. (2019), "BERT: Pre-training of deep bidirectional transformers for language understanding", presented at NAACL HLT 2019 – 2019 Conference of the North American Chapter of the Association for Computational Linguistics. [28]

Elloumi, Z. et al. (2018), "Analyzing learned representations of a deep ASR performance prediction model", in *Proceedings of the 2018 EMNLP Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP*, Association for Computational Linguistics, Brussels, https://www.aclweb.org/anthology/W18-5402/. [18]

Galibert, O. et al. (2014), "The ETAPE speech processing evaluation", in *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, European Language Resources Association, Reykjavik, https://www.aclweb.org/anthology/volumes/L14-1/. [13]

Hernández-Orallo, J. (2017), "Evaluation in artificial intelligence: From task-oriented to ability-oriented measurement", *Artificial Intelligence Review*, Vol. 48/3, pp. 398-447. [34]

Hernández-Orallo, J. (2017), *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*, Cambridge University Press, New York. [36]

Howard, J. and S. Ruder (2018), "Universal language model fine-tuning for text classification", *arXiv*, pp. 328-339, http://nlp.fast.ai/ulmfit. [26]

Huang, H. et al. (2007), "Characterizing unmanned system autonomy: Contextual autonomous capability and level of autonomy analyses", *Proceedings Volume 6561, Unmanned Systems Technology IX*, https://doi.org/10.1117/12.719894. [33]

Huckaby, J. and H. Christensen (2012), “A taxonomic framework for task modeling and knowledge transfer in manufacturing robotics”*, AAAI Workshop – Technical Report*, No. WS-12-06, Association for the Advancement of Artificial Intelligence, Paolo Alta, CA, http://dx.doi.org/www.aaai.org. [23]

Hughes, C. and T. Hughes (2019), “What metrics should we use to measure commercial AI?”, *AI Matters*, Vol. 5/2, pp. 41-45, https://doi.org/10.1145/3340470.3340479. [4]

Jiménez, J. et al. (2021), “Methodological aspects for cognitive architectures construction: A study and proposal”, *Artificial Intelligence Review*, Vol. 54/32133-2192, https://doi.org/10.1007/s10462-020-09901. [9]

Kotseruba, I. and J. Tsotsos (2016), “A review of 40 years of cognitive architecture research: Core cognitive abilities and practical applications”, *arXiv*, Vol. 08602, http://dx.doi.org/arXiv:1610.08602. [5]

Langley, P. (2006), “Cognitive architectures and general intelligent systems”, *AI Magazine*, Vol. 27/2, p. 33, https://doi.org/10.1609/aimag.v27i2.1878. [39]

Langley, P., J. Laird and S. Rogers (2009), “Cognitive architectures: Research issues and challenges”, *Cognitive Systems Research*, Vol. 10/2, pp. 141-160, https://doi.org/10.1016/j.cogsys.2006.07.004. [8]

Leuski, A. and D. Traum (2008), “A statistical approach for text processing in virtual humans”, https://www.researchgate.net/publication/228597921_A_statistical_approach_for_text_processing_in_virtual_humans. [7]

Mathet, Y. et al. (2012), “Manual corpus annotation: Giving meaning to the evaluation metrics”, in *Proceedings of COLING 2012: Posters*, The COLING 2012 Organizing Committee, Mumbai, https://www.aclweb.org/anthology/C12-2079. [37]

Moon, S., S. Kim and H. Wang (2014), “Multimodal transfer deep learning with applications in audio-visual recognition”, *arXiv*, Vol. 3121, http://arxiv.org/abs/1412.3121. [24]

Neubert, J. et al. (2015), “The assessment of 21st century skills in industrial and organizational psychology: Complex and collaborative problem solving”, *Industrial and Organizational Psychology*, Vol. 8/2, pp. 238-268, https://doi.org/10.1017/iop.2015.14. [35]

Oparin, I., J. Kahn and O. Galibert (2014), “First Maurdor 2013 evaluation campaign in scanned document image processing”, *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5090-5094, https://doi.org/10.1109/ICASSP.201. [17]

Peters, M. et al. (2018), “Deep contextualized word representation”, *arXiv*, Vol. 04365, http://allennlp.org/elmo. [27]

Primi, R. et al. (2016), “Mapping questionnaires: What do they measure?”, *Estudos de Psicologia (Campinas)*, Vol. 36/e180138., https://doi.org/10.1590/1982-0275201936e180138. [30]

Prokopalo, Y. et al. (2020), “Evaluation of lifelong learning systems”, in *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, Luxembourg, https://hal.archives-ouvertes.fr/hal-02496971. [12]

Razavian, A. et al. (2014), “CNN features off-the-shelf: An astounding baseline for recognition”, *arXiv*, Vol. 6382, http://dx.doi.org/arXiv:1403.6382. [25]

Shibata, K. (2017), "Functions that emerge through end-to-end reinforcement learning –The direction for artificial general intelligence", *arXiv*, Vol. 0239, http://dx.doi.org/arXiv:1703.02239. [22]

Shneier, M. et al. (2015), "Measuring and representing the performance of manufacturing assembly robots"*, NIST Interagency/Internal Report (NISTIR)*, No. 8090, National Institute of Standards and Technology, Gaithersburg, MD, https://doi.org/10.6028/NIST.IR.8090. [31]

Stasse, O. et al. (2018), "Benchmarking the HRP-2 humanoid robot during locomotion", *Frontiers Robotics AI, 5(NOV)* 8 November, https://doi.org/10.3389/frobt.2018.00122. [21]

Tolan, S. et al. (2020), "Measuring the occupational impact of AI: Tasks, cognitive abilities and AI benchmarks", *Journal of Artificial Intelligence Research*, Vol. 71, https://doi.org/10.1613/jair.1.12647. [32]

Ye, P., T. Wang and F. Wang (2018), "A survey of cognitive architectures in the past 20 years", *IEEE Transactions on Cybernetics*, Vol. 48/12, pp. 3280-3290, https://doi.org/10.1109/TCYB.2018.2857704. [6]

Notes

[1] http://rockinrobotchallenge.eu/

[2] www.quaero.org/

[3] www.eu-robotics.net/eurobotics/about/projects/rockeu2.html

[4] https://sciroc.org/

[5] http://challenge-rose.fr/

[6] https://metricsproject.eu/

[7] https://bicasociety.org/cogarch/

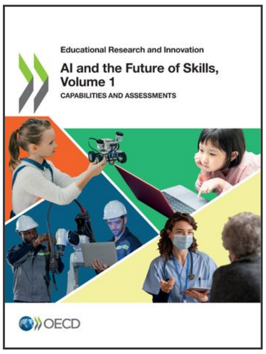
[8] https://web.archive.org/web/20100315140823/http://ai.eecs.umich.edu/cogarch0/common/capa.html

[9] The usefulness of having cognitive architectures to produce general artificial intelligence is presented in Langley (2006[39]).

[10] www.onetonline.org/

[11] Learning data is also often subject to human annotation, which can be related to the concept of *Fauxtomation*.

[12] https://ec.europa.eu/growth/single-market/goods/building-blocks/conformity-assessment_en



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:

Avrin, Guillaume (2021), “Assessing artificial intelligence capabilities”, in OECD, *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/47d04fe3-en

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

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Extracts from publications may be subject to additional disclaimers, which are set out in the complete version of the publication, available at the link provided.

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