

7 AI direct tests: LNE and NIST evaluations

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

Guillaume Avrin, Laboratoire National de Métrologie et d'Essais

Swen Ribeiro, Laboratoire National de Métrologie et d'Essais

Edited by: Abel Baret, OECD

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

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

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

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

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

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

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

The need for systematising AI and robotics evaluations

Not all tasks automated by AI and robotics have been evaluated

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

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

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

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

Not all AI tasks are relevant for humans

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

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

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

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

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

Framework structure

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

An evaluation consists in:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Comparison with human capabilities

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

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

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

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

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

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

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

Evaluation campaigns of AI capabilities

Evaluation campaigns at LNE, NIST and other institutions

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

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

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

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

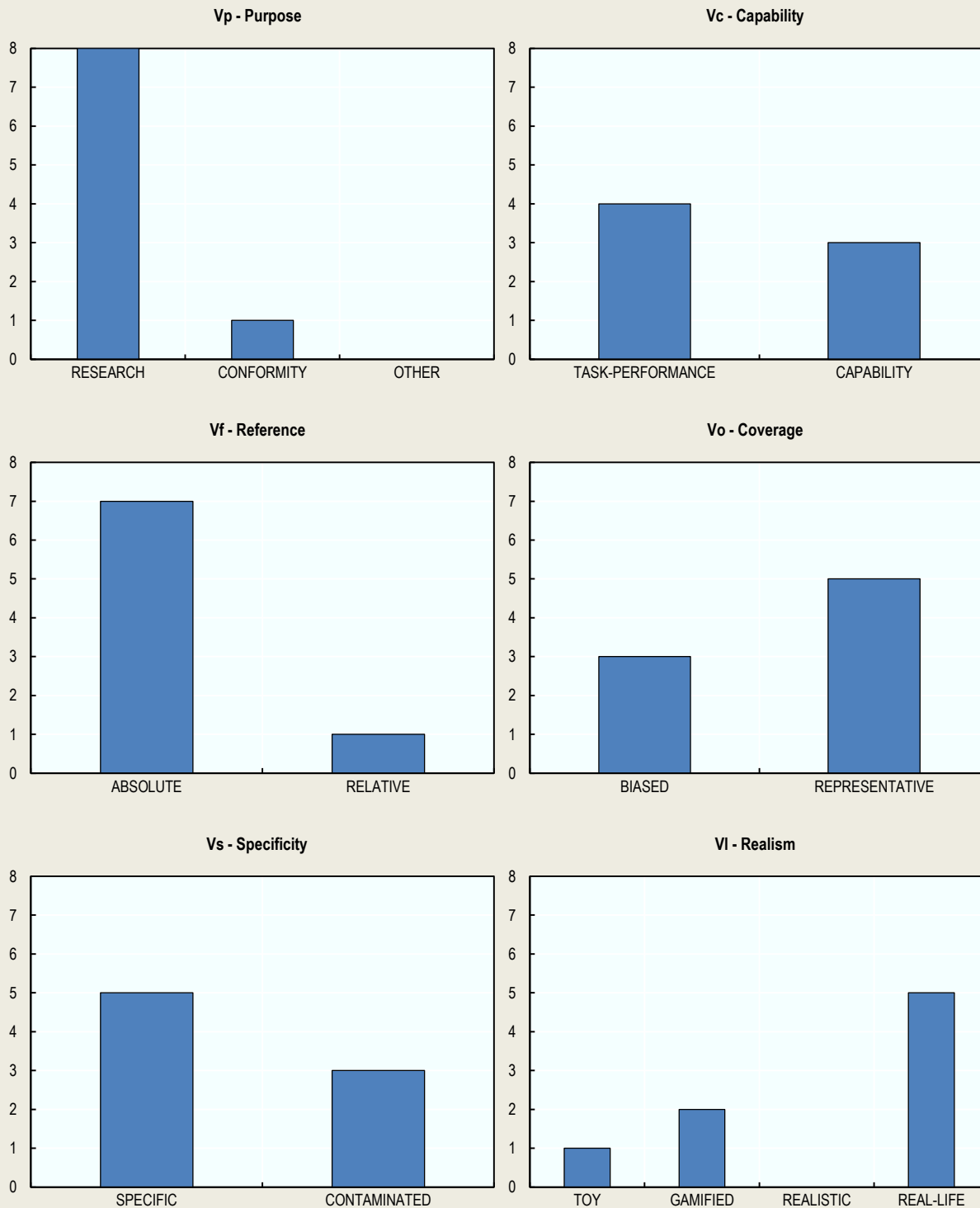
With regards to the **Validity** group (“Does it measure what we want to measure?”) and illustrated by the first six graphs, five of six facets (i.e. Purpose, Capability, Reference, Coverage and Specificity) have frequencies similar to the 36 benchmarks from the previous chapter (see Figure 7.1). The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

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

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

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

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




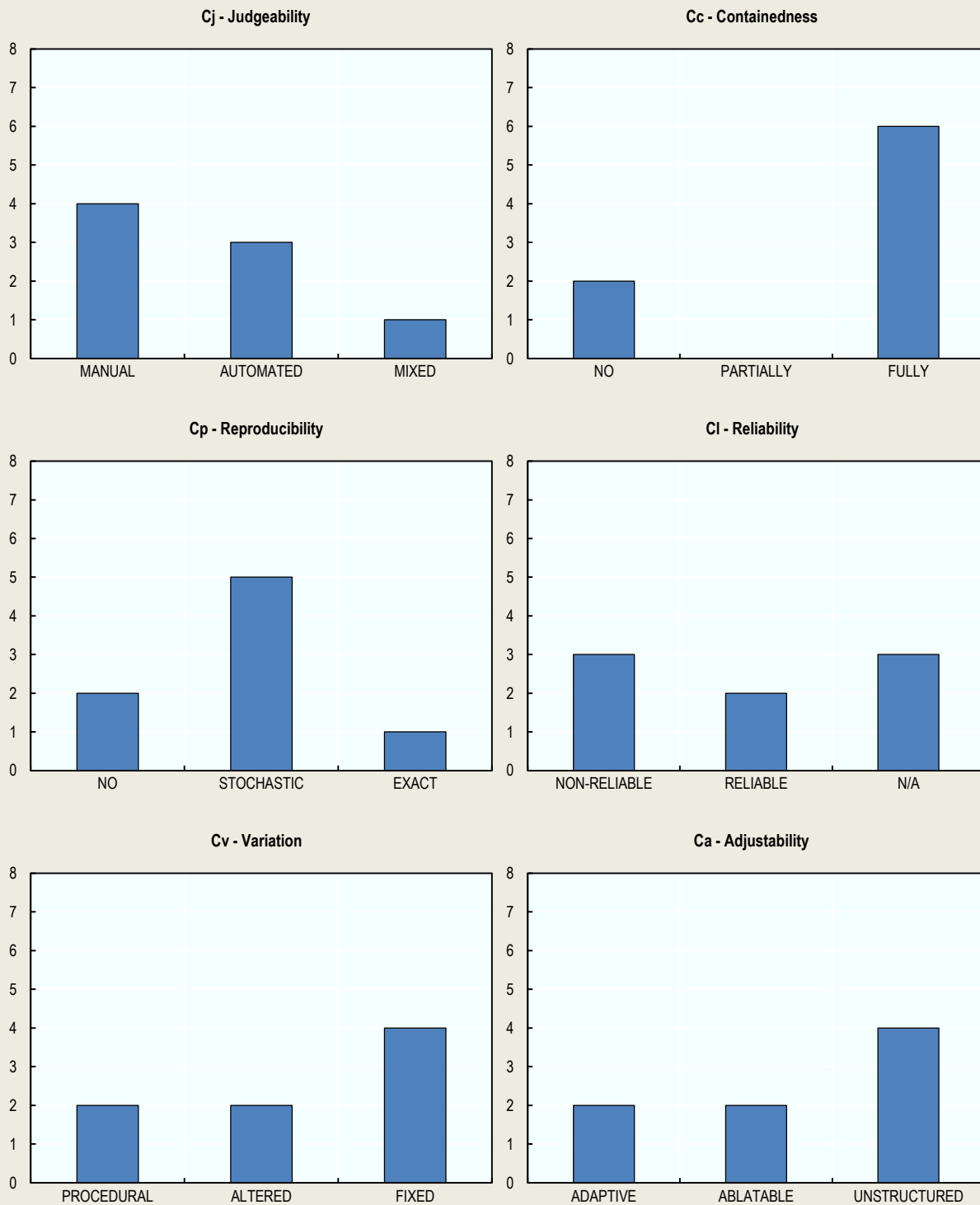
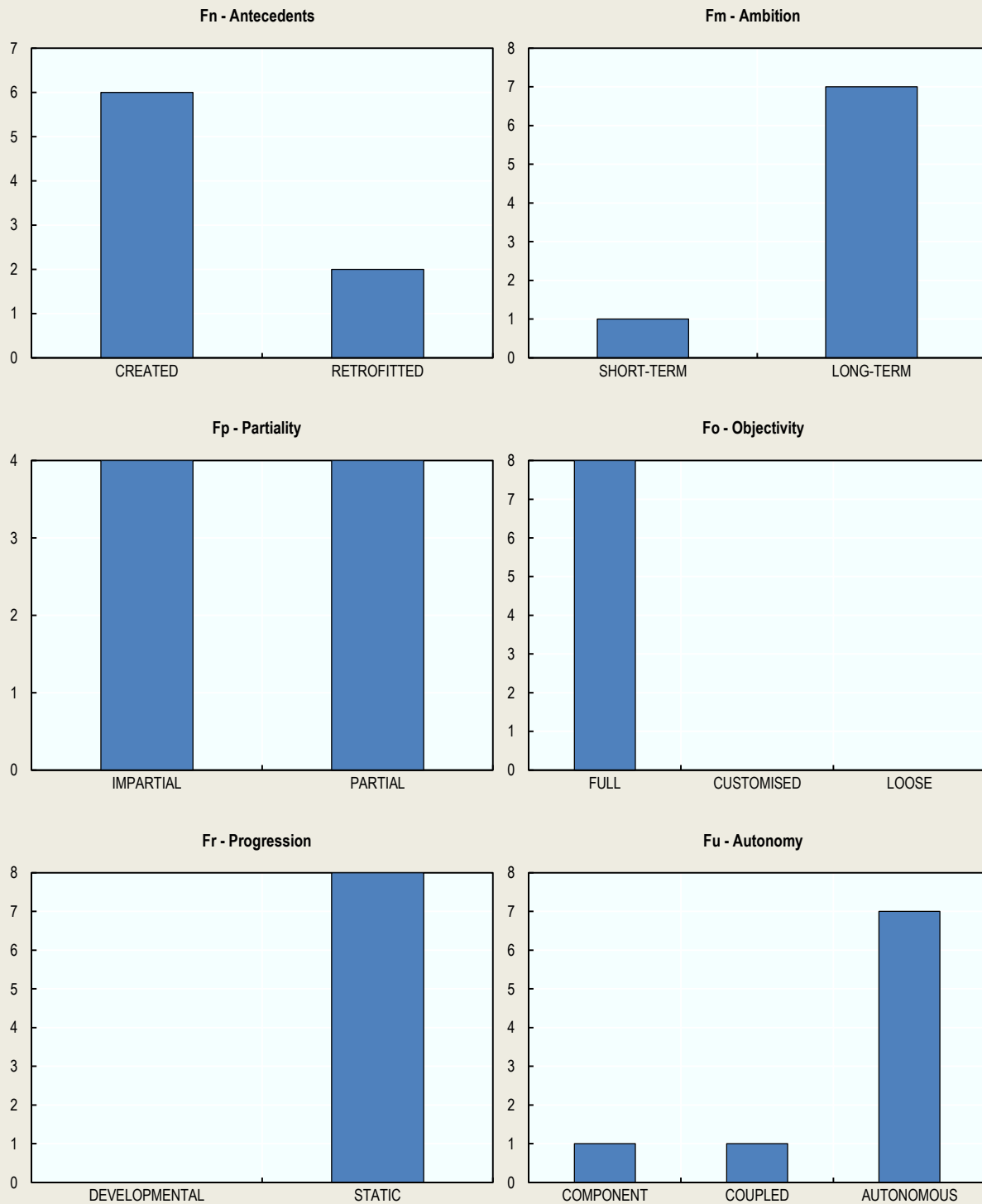
StatLink  <https://stat.link/snb480>

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



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

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



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

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

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

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

Natural language processing

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

Text processing and text comprehension

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

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

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

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

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

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

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

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

Speech processing

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

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

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

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

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

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

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

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

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

Computer vision

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

Recognition

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

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

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

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

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

Motion analysis

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

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

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

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

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

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

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

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

Robotics

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

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

Locomotion

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

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

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

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

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

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

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

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

Manipulation

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

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

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

Table 7.6. Manipulation task examples and associated evaluation campaigns

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

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

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

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

Going beyond “basic” interaction

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

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

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

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

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

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

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

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

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

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

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

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

Limitations and uncovered tasks from AI evaluations

Uncharted tasks

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

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

Other AI capabilities

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

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

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

Conclusion

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

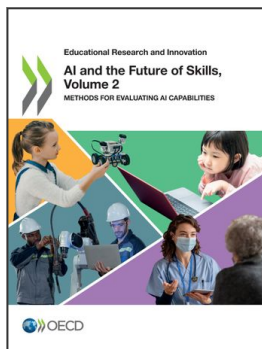
Notes

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

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

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

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



From:
AI and the Future of Skills, Volume 2
Methods for Evaluating AI Capabilities

Access the complete publication at:
<https://doi.org/10.1787/a9fe53cb-en>

Please cite this chapter as:

Messina, Elena, Guillaume Avrin and Swen Ribeiro (2023), "AI direct tests: LNE and NIST evaluations", in OECD, *AI and the Future of Skills, Volume 2: Methods for Evaluating AI Capabilities*, OECD Publishing, Paris.

DOI: <https://doi.org/10.1787/982e32ab-en>

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>.