

6

A framework for characterising evaluation instruments of AI performance

Anthony G Cohn, University of Leeds

José Hernández-Orallo, Universitat Politècnica de València

Edited by: Sam Mitchell and Stuart Elliot, OECD

This chapter presents and discusses an approach to categorising benchmarks, competitions and datasets, jointly referred to as evaluation instruments of artificial intelligence (AI) performance. It proposes a set of 18 facets to distinguish and evaluate existing and new evaluation instruments, rating a sample of 36 evaluation instruments according to these facets. With a rubric composed of these 18 facets, four raters evaluate the sample, illustrating how well facets help analyse aspects of AI appraised by each evaluation instrument. In this way, the chapter proposes a framework that the OECD and third parties (researchers, policy makers, students, etc.) can use to analyse existing and new evaluation instruments.

Several studies focus on numeric comparison and the evolution of performance for a range of evaluation instruments of artificial intelligence (AI) (Martínez-Plumed et al., 2021^[1]; Ott et al., 2022^[2]). However, these studies only track the evolution of the progress of AI systems themselves. As such, they do not provide insight into how evaluation instruments such as benchmarks, competitions, standards and tests are also evolving. Nor do they indicate whether the measures are meeting the demands of a more comprehensive evaluation beyond some simple metrics. In response to this gap in the AI evaluation field, this chapter proposes a methodology to characterise the AI evaluation landscape. It will also assess the extent to which evaluation instruments can be used to evaluate the capabilities of AI systems over time.

There are thousands of evaluation instruments across all areas of AI, which makes it challenging to characterise the landscape of AI evaluation. As AI techniques evolve, they are also increasingly complex and diverse. Because of this, it is hard to analyse this evaluation landscape in a meaningful way. As a first step to overcome these challenges, this chapter presents and discusses an approach to categorising AI evaluation instruments. This categorisation is performed with a set of 18 facets, which are proposed to distinguish and evaluate the characteristics of existing and emerging evaluation instruments.

This chapter codes a sample of 36 evaluation instruments to evaluate how well the facets work in general and to what extent they help map the landscape of evaluation instruments and distinguish their differences. An evaluation instrument classification based on these facets may inform the design of future evaluation instruments. It is not clear if a single universal evaluation instrument will ever be feasible, or even a battery for each domain (vision, reasoning, etc.). Certainly, that ideal has eluded the community so far. The chapter aims to help direct future efforts in the evaluation of AI systems rather than find facet values that are valid for all evaluation instruments.

The 36 evaluation instruments classified in this chapter are only a cross-section of the thousands across all fields of AI research. Beyond the insights extracted from the sample, this paper and the rubric developed for the facets should serve as a reference for third parties (e.g. other researchers) to analyse other existing and newly proposed evaluation instruments. The work demonstrates that a set of evaluation instruments can be coded according to the facets in a relatively reliable manner. The resulting values reveal some interesting patterns about the characteristics of evaluation instruments used in the field.

The rest of the chapter is organised as follows. The second section presents the proposed 18 facets and a rubric that explains how facet values should be chosen. Then the criteria for selecting the 36 evaluation instruments and the methodology the raters used to apply the rubric is presented. The next section discusses the level of disagreement between raters for each facet and evaluation instrument. The penultimate section analyses the ratings of the 36 evaluation instruments, and what they reveal about this group of evaluation instruments. Finally, findings and possible future work is discussed in the final section.

Characterising AI evaluation instruments

The project initially hoped to find and build on existing methods to characterise evaluation instruments, but at the start of the project it became apparent a methodology that could be applied consistently across the AI evaluation field did not yet exist. Therefore, the project has defined a novel framework for this task inspired by work outside of AI research that has developed more systematic coverage of evaluation methods: the new set of facets proposed to evaluate evaluation instruments are inspired by psychological testing. The terminology used in this chapter is based on common use in AI, but also incorporates terms and concepts from the Standards for Educational and Psychological Testing by the American Educational Research Association (AERA, APA, NCME, 2014^[3]).

The following list proposes 18 facets to characterise existing and future evaluation instruments for AI. Each facet is followed by the values according to which an evaluation instrument can be classified in brackets. Some values indicate “(specify)”, which means the rater must respond in free text for that value. The colour

blue indicates cases where a facet has a preferred value, *in general* (for some evaluation instruments, another value may be preferred). However, some facets do not have a preferred value and therefore all values are left in black.

The facets are grouped into three main categories following the three main groups given by the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014^[3]): Validity, Consistency and Fairness. These groups deal with *what AI performance is measured*, *how it is measured* and *what AI system is measured*, respectively.

Validity facets (Does it measure what it should?)

- **Capability** [TASK-PERFORMANCE (specify), CAPABILITY (specify)]: does the evaluation instrument just measure observed (aggregated) performance on a TASK (e.g. protein folding, credit scoring) or can the evaluation instrument also measure a CAPABILITY (e.g. object permanence, dealing with negation)?
- **Coverage** [BIASED (specify), REPRESENTATIVE]: does the evaluation instrument cover a BIASED or unbiased (REPRESENTATIVE) distribution of what is meant to be measured?
- **Purpose** [RESEARCH, CONFORMITY, OTHER (specify)]: is the benchmark meant to foster research or development, or to certify whether an AI system conforms with some level or standard?
- **Realism** [TOY, GAMIFIED, REALISTIC, REAL-LIFE¹]: to what extent is the evaluation instrument a toy problem or a complex gamified problem? Is it a realistic setting (e.g. a simulated scenario, a lab or testing facility) or is the evaluation itself happening in real life?
- **Reference** [ABSOLUTE, RELATIVE (specify)]: are results reported as an absolute metric (criterion-referenced) or are they reported as a relative (percentage) metric to a reference (norm referenced), e.g. human performance?
- **Specificity** [SPECIFIC, CONTAMINATED]: are the results precisely aligned with what is meant to be measured or contaminated by other skills or tasks?

Consistency facets (Does it measure it effectively and verifiably?)

- **Adjustability** [UNSTRUCTURED, ABLATABLE², ADAPTIVE]: is the analysis of results on the set of instances unstructured?; has the evaluation instrument identified a set of meta-features such as difficulty or dimension that could be used to analyse the results by these dimensions (ablatable)?; or are these meta-features used to adaptively or adversarially choose the instances to test more informatively (adaptive)?
- **Containedness** [FULLY-CONTAINED, PARTIAL-INTERFERENCE (specify), NOT-CONTAINED (specify)]: Once started, is the testing isolated from external factors or interference possibly affecting results (human participants, online data, weather, etc.)?; is there some partial interference not affecting the results significantly?; or is it dependent on external resources and conditions?
- **Judgeability** [MANUAL, AUTOMATED, MIXED]: is scoring manual (e.g. through human questionnaires or judges) or automated (e.g. correct answers or optimality function) or a mixture?
- **Reliability** [RELIABLE, NON-RELIABLE, N/A]: does the evaluation present sufficient repetitions, episode length or number of instances to give low variance for the same subject when applied again (test-retest reliability)? If the testing methodology or the common use of the evaluation instrument is not clear, then N/A may be the most appropriate facet value.
- **Reproducibility** [NON-REPRODUCIBLE, STOCHASTIC, EXACT]: is the evaluation non-reproducible, with results biased or spoiled if repeated?; does the evaluation instrument have stochastic components leading to different interactions?; or are the results completely reproducible,

i.e. can the same exact test (inputs, interaction, etc.) be generated again for another (or the same) competitor?

- **Variation** [FIXED, ALTERED, PROCEDURAL]: is the evaluation based on fixed datasets?; have the instances been altered by adding post-processing variations (noise, rotations, etc.)?; or have the instances been created (e.g. using procedural generation³)?

Fairness facets (Does it treat all test takers equally?)

- **Ambition** [SHORT, LONG]: when the evaluation instrument was created, was it aiming at the short term (improving on the state of the art) or long term (more ambitious goals)?
- **Antecedents** [CREATED, RETROFITTED (specify)]: is it devised purposely for AI or adapted from tests designed to test humans?
- **Autonomy** [AUTONOMOUS, COUPLED (specify), COMPONENT]: is it measuring an autonomous system, coupled with other systems (e.g. humans) or as an isolated component?
- **Objectivity** [LOOSE, CUSTOMISED, FULLY-INDEPENDENT]: is it loosely defined, customised to each participant or does the evaluation instrument have a predetermined independent specification?⁴
- **Partiality** [PARTIAL (specify), IMPARTIAL]: does the evaluation instrument favour particular technologies, conditions or cultures that should not have an influence on the result of the evaluation?⁵
- **Progression** [STATIC, DEVELOPMENTAL]: Is the score measuring a capability at one moment or is it evaluating the development of the capability of the system within the test?

Facets with preferred values reflect suggestions about directions for changing the characteristics of evaluation instruments to improve them. For example, researchers testing AI should prefer an evaluation instrument that is RELIABLE (**Reliability**) to an evaluation instrument that is NON-RELIABLE, all things being equal. Facets that do not have preferred values are useful for categorising evaluation instruments in terms of other characteristics that may be useful for a particular purpose. For example, if a researcher is using an evaluation instrument that measures TASK-PERFORMANCE (**Capability**), then they cannot draw conclusions about that AI's capabilities based on its performance on that evaluation instrument alone. An evaluation instrument that measures CAPABILITY, however, could be used to draw such conclusions.

Some of the facets, including across groups, are also closely related, such as {**Variation, Adjustability, Coverage**} or {**Objectivity, Reproducibility**}. One would expect that an evaluation instrument with a FULLY-INDEPENDENT value for **Objectivity** is more likely to be rated as EXACT for **Reproducibility**, for example.

Finally, the variability of measurement is also an important concept when evaluating evaluation instruments. In other words, how many changes can be made to an evaluation instrument for each AI system evaluation before the different evaluation results are no longer comparable? The term *accommodation* is “used to denote changes with which the comparability of scores is retained, and the term *modification* is used to denote changes that affect the construct measured by the test” (AERA, APA, NCME, 2014^[3]). This is important for **Specificity, Variation, Objectivity** and **Containedness**, as it indicates whether accommodations of the same test could evaluate different AI systems and even humans in a comparable way.

Evaluation instrument selection and rating methodology

Evaluation instrument selection

Evaluation instruments that met the following criteria were considered for inclusion:

- *Potential interest to understand the future of AI skills*: an evaluation instrument might be considered interesting if high AI performance can be regarded as indicating a noteworthy change in the capabilities of AI in general. In other words, progress in this evaluation instrument requires significant enhancement of AI techniques beyond the specific requirements of the evaluation instrument.
- *Diversity in the kind of task*: the evaluation instrument sample should cover a variety of domains (vision, natural language, etc.), formats (competitions, datasets, etc.) and types of problems (supervised/unsupervised learning, planning, etc.).
- *Popularity*: how many teams have already used this evaluation instrument? How many published papers refer to it? More popular evaluation instruments were preferred in the selection. Citations to the original papers introducing the evaluation instrument, the number of results on websites such as paperswithcode.com, etc., can be used as proxies to evaluate popularity. The possibility of industry-related evaluation instruments being less popular than research-oriented evaluation instruments was also considered.
- *Currency*: evaluation instruments still in active use or recently introduced were preferred rather than those that have fallen out of use.

The source of the evaluation instruments was mostly repositories⁶ and surveys, institutions such as National Institute of Standards and Technology and Laboratoire National de Métrologie et d'Essais, and competitions at AI conferences. The study then considered possible gaps and overlaps in the sample's coverage of domains. At the time of selection, only a rough estimate of potential preferred categories was possible for each evaluation instrument. The evaluation instruments have been categorised into six AI domains; the total count is more than 36 as multiple evaluation instruments tested AIs on more than one domain. For example, the Bring-Me-A-Spoon evaluation instrument (Anderson, 2018_[4]) evaluates AI on language understanding and robotic performance. The complete list of 36 selected evaluation instruments with their descriptions are shown in Annex Table 6.A.1.

Table 6.1. Primary testing domain of sampled evaluation instruments

AI domain	Reasoning	Language	Robotics	Vision	Video games*	Social-emotional
Number of evaluation instruments	12	11	7	5	6	1

Note: *Given the wide diversity of inputs across different video games and that different tasks within the same video game can require different capabilities, evaluation instruments based on video games were categorised separately.

These evaluation instruments cover a good distribution of benchmarks, competitions and datasets, although some can be considered to be in two categories. The term “test” to refer to an evaluation instrument is less usual. About half of the 36 evaluation instruments require use of language in the inputs and/or outputs, while about half require some kind of perception (mostly computer vision). There is some overlap in these two groups. Only a few evaluation instruments are related to navigation and robotics, in virtual (e.g. video games) or physical environments. A small number are related to more abstract capabilities or problems related to planning or optimisation.

Table 6.2. Type of sampled evaluation instruments

Evaluation instrument type	Competition	Benchmark	Dataset
Number of evaluation instruments	20	12	10

Note: Some evaluation instruments are a combination of types.

Rating methodology

A protocol to refine the rubric and to cover as many evaluation instruments as possible with available resources, is explained below. This protocol can be adapted to other situations or incorporate ideas from consensus-based ratings or the Delphi method (Hsu and Sandford, 2007^[5]). First, Anthony Cohn and José Hernández-Orallo acted as co-ordinators for the rating process, choosing four raters – Julius Sechang Mboli, Yael Moros-Daval, Zhiliang Xiang and Lexin Zhou (Cohn et al., 2022^[6]).⁷ Raters were AI-related undergraduate and graduate students and were recruited through a selection process, including interviews. Once the raters were appointed, each rater was given some meta-information about each evaluation instrument (acronym, name, major sources, what it measures, etc.) and completed other general information about each evaluation instrument (see Annex Table 6.A.1). They were also asked some information about their own completion, such as time taken (in hours). In all, 36 evaluation instruments were evaluated in this manner.

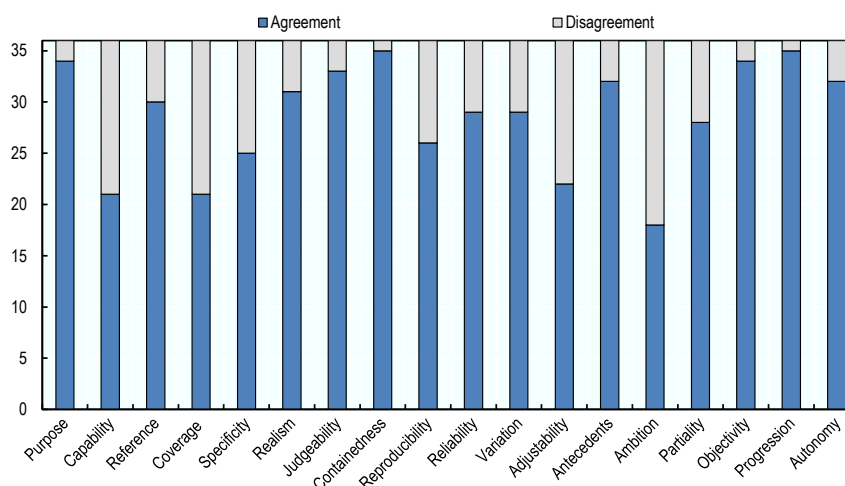
The evaluation instruments were rated in four batches. The first two evaluation instruments (Batch 1) were used by all co-ordinators and raters. All four raters then rated the next 11 evaluation instruments (Batch 2) and held discussions to refine the rubric. After observing consistent ratings across all four raters, only two raters rated each of the next ten evaluation instruments (Batch 3) and the final set (Batch 4) as this allowed a higher number of evaluation instrument evaluations with the same level of resources. Raters worked independently but discussed ratings after Batches 3 and 4, leading to some rating changes after the discussion. Annex Table 6.A.1 gives an overview of all 36 evaluation instruments and the batches they were evaluated in.

Analysis of rater consistency


The pattern of agreement or disagreement among the raters tends to vary depending on factors such as facet complexity, available information on the evaluation instrument and so on (see Figure 6.1). The most notable observations were the following:

- There is *consistent* agreement on **Progression, Autonomy, Purpose, Judgeability, Containedness, Objectivity** and **Autonomy** across all batches.
- There is *moderate* agreement on **Reference, Realism, Reproducibility, Variation** and **Partiality**. Notably, **Realism** has the largest number of values, but still obtains agreement well across evaluation instruments.
- There is the *least* agreement on **Capability, Coverage, Specificity, Adjustability** and **Ambition**, facets with mostly with binary options, with disagreement ranging from a third to a half of the evaluation instruments.

Figure 6.1. Rater agreement across all facets



Note: Agreements on facet value ratings for the 36 direct measures. “Agreement” means unanimous agreement and “Disagreement” covers all other cases.

StatLink  <https://stat.link/zk0q2p>

Overall, the results suggest that facets can be coded relatively reliably. Two factors help explain the lower rater consistency for some facets:

- To make justifiable decisions for facets like **Coverage** and **Specificity**, raters often needed to seek related literature for support when the answers were not clear from the specifications of evaluation instruments. Whether an evaluation instrument is specific (**Specificity**) and general (**Coverage**) enough for the measuring of certain capabilities is indeed hard to judge depending solely on the specifications. Furthermore, information extracted from different sources might lead to disagreements on selections.
- The subjectivity of a facet could also contribute to value divergences. This might be a reasonable explanation for inconsistent selections in **Capability**, **Adjustability** and **Ambition** since they allow raters more space for subjective interpretations. While relevant information regarding **Capability** and **Ambition** is often stated in the evaluation instrument specifications, these statements can somehow be interpreted in different degrees or ways. For example, an evaluation instrument for natural language understanding (NLU) could aim at improving state-of-the-art performance (short term) or measuring agents’ capabilities regarding NLU (long term); object recognition could be argued as a visual capability or a specific task.

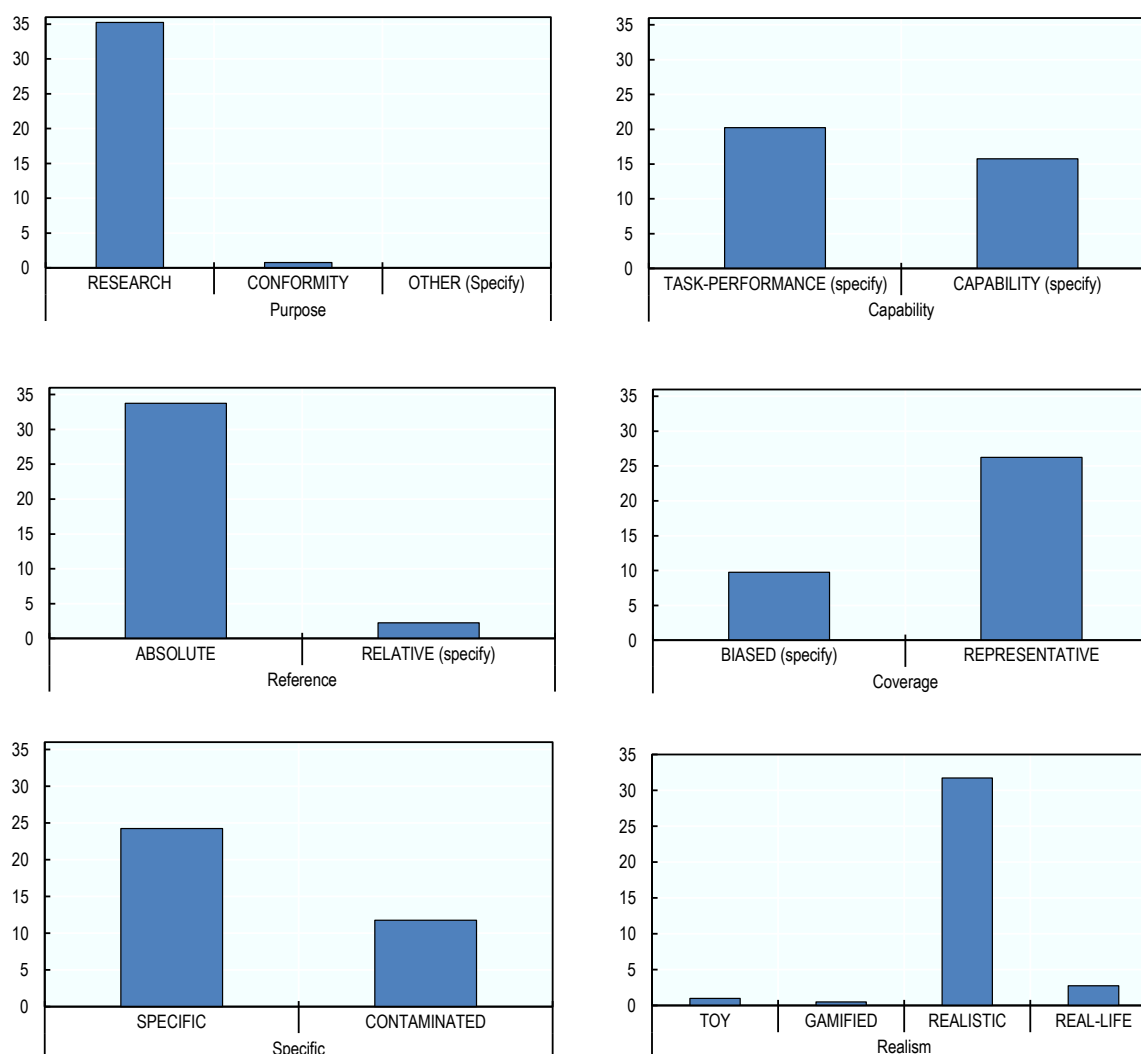
Analysis of facet values

Validity facets (Does it measure what it should measure?)

The findings indicate that sampled evaluation instruments are primarily designed for academic research, use absolute metrics and are divided roughly equally between measuring capabilities and specific performance tasks (see Figure 6.2).

- Nearly all the chosen evaluation instruments are aimed at promoting RESEARCH (**Purpose**) and predominantly use ABSOLUTE metrics (**Reference**).

Figure 6.2. Rater value selection on validity facets



StatLink  <https://stat.link/y0h5wp>

- The distribution of evaluation instruments measuring a specific task and those aiming for capability assessment is nearly balanced (**Capability**). This points to an ongoing debate in the field about the focal point of evaluation – performance or capabilities.
- Most evaluation instruments were classified as REPRESENTATIVE (**Coverage**). However, around 27% of the evaluation instruments are BIASED.
- About two-thirds of the evaluation instruments were SPECIFIC (**Specificity**). The remaining one-third were classified as CONTAMINATED, meaning the results may not fully align with the intended measurement objectives.
- Approximately 80% of evaluation instruments are REALISTIC (**Realism**), showing a strong inclination towards solving practical problems. Nonetheless, most evaluations have yet to be conducted in real-world settings.

The analysis found these areas for improvement:

- Most of the selected evaluation instruments that measure a capability (**Capability**) do not necessarily measure it reliably.

- Representativeness in the current evaluation instruments (**Coverage**) remains limited.
- Conducting more evaluations in real-world settings would promote development of more effective AI systems (**Realism**).

Consistency facets (Does it measure it effectively and verifiably?)

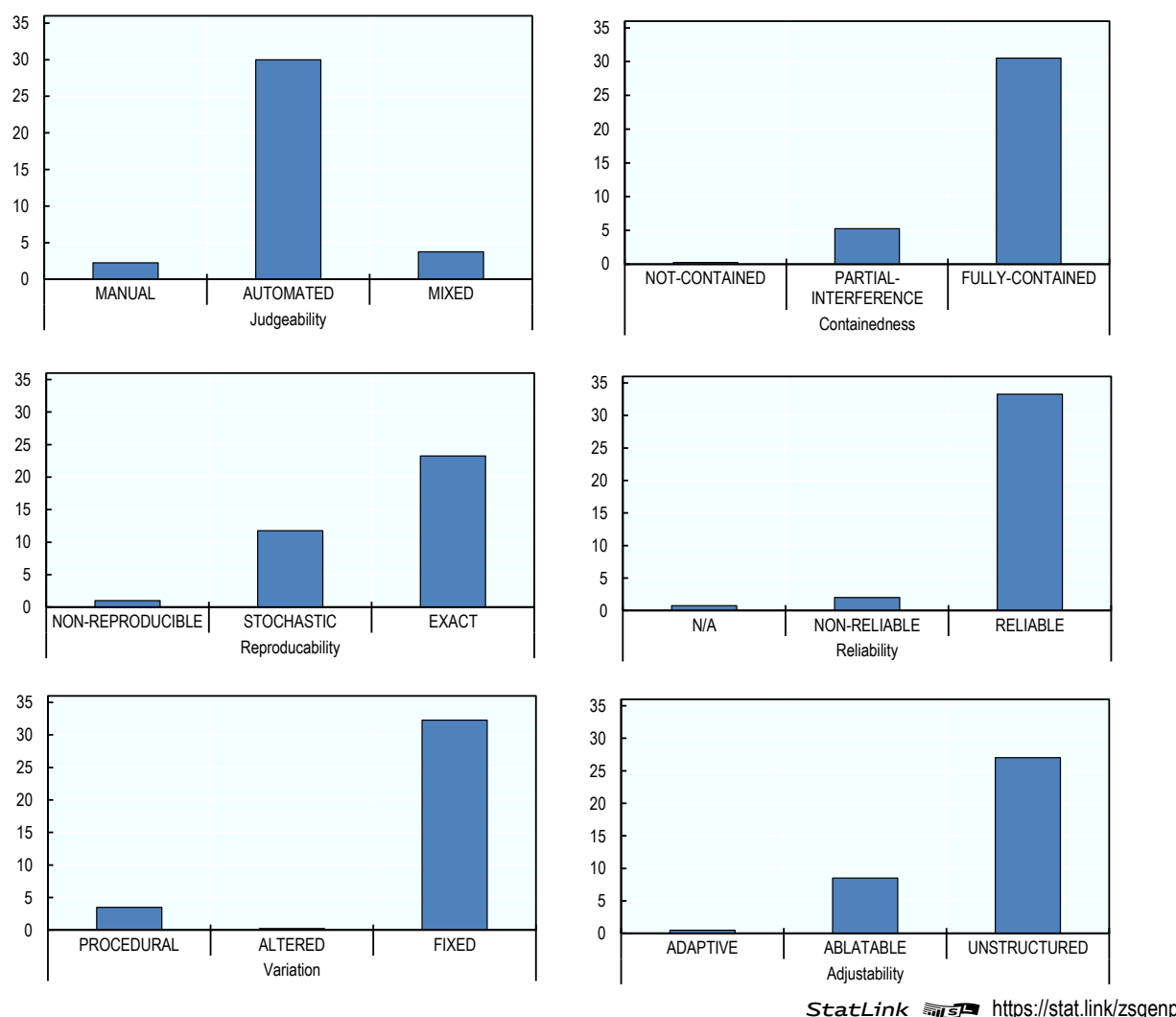
The results indicate that sampled evaluation instruments are mostly independent from external factors, are reliable, use fixed datasets and allow automated scoring (see Figure 6.3):

- Nearly all the selected evaluation instruments fall under the FULLY-CONTAINED category (**Containedness**), suggesting a high level of independence from external factors during assessments. This is a desirable feature for maintaining the integrity of an evaluation.
- Most evaluation instruments are classified as RELIABLE (**Reliability**), which lends credibility to the evaluation process.
- When it comes to **Judgeability**, most evaluation instruments employ AUTOMATED scoring instead of MANUAL or MIXED. While automated scoring generally offers more objectivity and speed, it does raise questions about the definition of the scoring metrics. For instance, determining the quality of a robotic dancer or cook through automated means can be challenging.
- In terms of **Variation**, nearly all evaluation instruments rely on FIXED datasets, which could limit the diversity in evaluation methods. For example, adding noise to the data could provide insights into the model's robustness.
- Most evaluation instruments are either UNSTRUCTURED or ABLATABLE (**Adjustability**), with few being ADAPTIVE. The absence of adaptive tests could be attributed to their operational complexity.

Further improvement is recommended in the following areas:

- introducing more diversity in the evaluation process, perhaps by adding post-processing variations or developing methods to cover intrinsic variations.
- encouraging more adaptive testing methods to evaluate how systems adapt to varying levels of difficulty.

Figure 6.3. Raters value selection on consistency facets



Fairness facets (Does it treat all test takers equally?)

The results show that most sampled evaluation instruments are impartial, objective and focus on static performance and AI systems working in isolation (see Figure 6.4):

- The raters found IMPARTIAL evaluation instruments account for 90% of the data (**Partiality**). However, the actual value might be lower since it is often hard to detect impartiality in the information given about an evaluation instrument. For instance, in an evaluation instrument for benchmarking clinical decision support systems, the training set may only include Latin American patients. However, the test set may include international patients.
- Virtually all the analysed evaluation instruments are classified as FULLY-INDEPENDENT (**Objectivity**), which favours fairness in evaluation.
- Nearly all evaluation instruments evaluate the AI systems are STATIC as opposed to DEVELOPMENTAL (**Progression**). This is possibly because many applications consider final performance as more important than how the system's performance evolves over time. It is also much harder to systematically evaluate AI systems over time. However, DEVELOPMENTAL evaluation instruments could give more insights into how the models learn with different data;

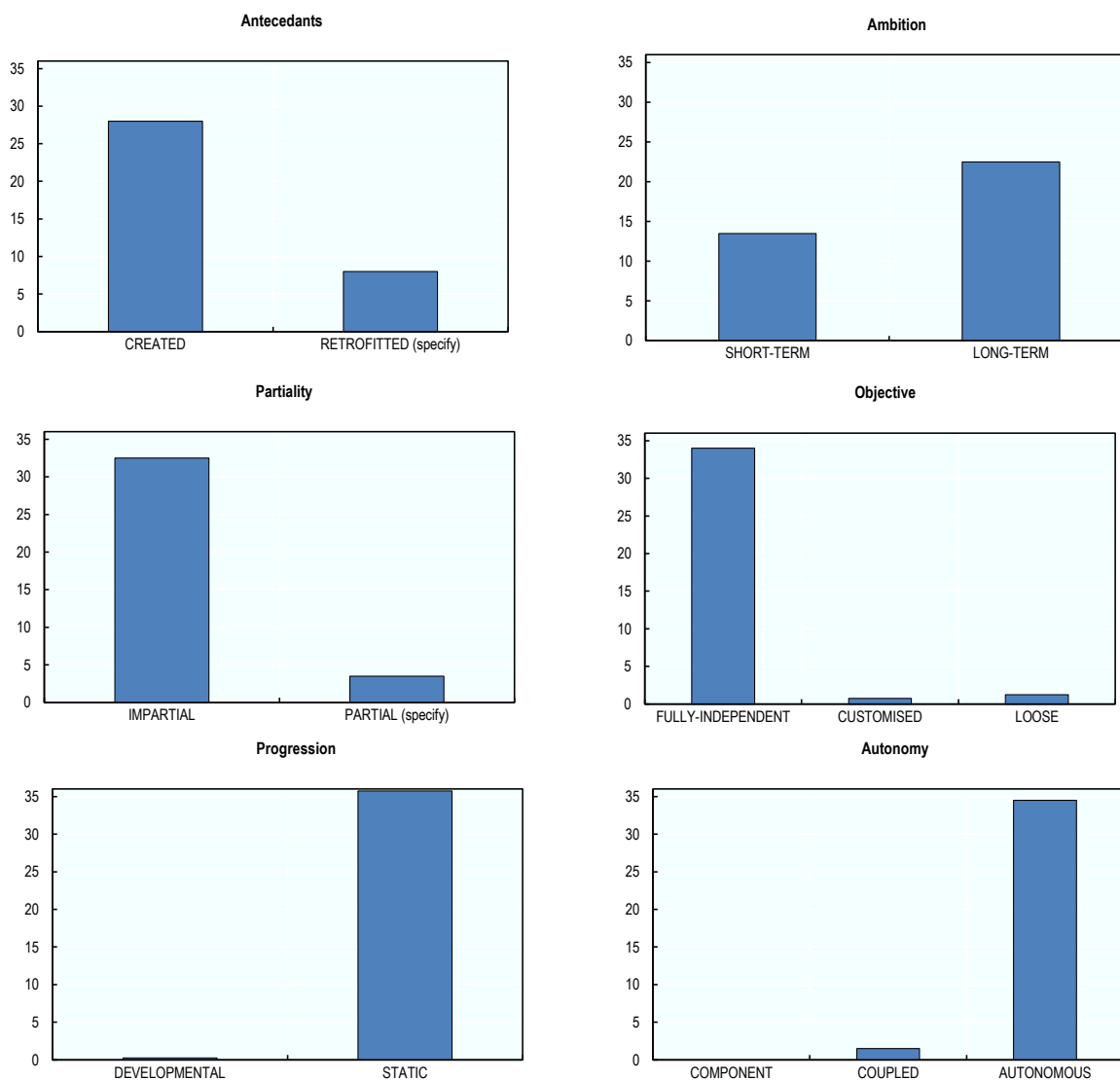
detect when and why things go wrong during the training phase; and identify the trade-off between training data size, time and performance.

- Only one of 36 evaluation instruments measured the performance of AI systems working with humans (COUPLED), rather than working autonomously or as an isolated component.

More efforts are needed to develop benchmarks that evaluate AI performance:

- at intervals through time as AI systems continue to develop (DEVELOPMENTAL).
- for those AIs that work together with humans (COUPLED).

Figure 6.4. Raters value selection on fairness facets



StatLink  <https://stat.link/e6jmcb>

Observations across facet groups

When looking at the distribution of facet values per evaluation instrument at a global level, evaluation instruments related to robotics and the physical world (Robocup, 2023^[7]; Robocup@Home, 2023^[8];

Lifelong Robotic Vision, 2023^[9]) have more variability in several respects. Many of them are judged manually as opposed to being automatically scored (**Judgeability**). Many have measures that are more realistic or close to real life (**Realism**). Testing is not always isolated from external factors but is often partially influenced by them (**Containedness**). In addition, they do not always measure systems autonomously, but sometimes with human interactions (**Autonomy**). One of the most popular evaluation instruments in the history of AI, ImageNet (Deng et al., 2010^[10]), is the only one which more than half of raters found to be partial (PARTIALITY), and the only evaluation instrument continuously rated as biased in **Coverage** (along with LibriSpeech). The disagreement in partiality may suggest that some sources of partiality are only discovered after the repeated use of an evaluation instrument and not identified by everyone immediately.

General Video Game Artificial Intelligence (Perez-Liebana et al., 2019^[11]) is a unique evaluation instrument that capitalises on the ablatable nature of video games, which can be altered easily by several characteristics or difficulty of the game. This is also going towards being procedural, but only to a limited extent as suggested by raters' values.

Finally, those evaluation instruments related to natural language, and especially WSC (Levesque, Davis and Morgenstern, 2011^[12]), GLUE (Wang et al., 2018^[13]), SUPERGLUE (Wang et al., 2019^[14]), Physical IQa (Bisk et al., 2020^[15]), SocialQA (Sap et al., 2019^[16]), SQUAD2.0 (Rajpurkar et al., 2016^[17]), WikiQA (Yang, Yih and Meek, 2015^[18]) and sW/AG (Zellers et al., 2018^[19]), have high degrees of contamination in the **Specific** facet. This might reflect the difficulty of isolating capabilities when using natural language, as some basic natural language competency requires many other things. This is reflected by the success of language models recently doing a variety of tasks (Devlin et al., 2018^[20]; Brown et al., 2020^[21]; Hendrycks et al., 2021^[22]; Bommasani et al., 2021^[23]), since mastering natural language seems to be contaminated by so many other capabilities and skills.

Conclusion

The framework presented in this chapter aims to provide a foundation from which evaluation instruments can be systematically evaluated and their evolution tracked.

The proposed set of facets and associated rubric, as well as the results of the study of 36 evaluation instruments reported in this paper, can be useful for three different kinds of users in slightly different ways.

1. First, evaluation instrument creators can see what design choices in their evaluation instrument to modify from a first evaluation of its facets and see how it compares to other evaluation instruments.
2. Second, AI system developers can choose the most appropriate evaluation instruments according to the facet values, and better understand what to expect from the evaluation and what it means exactly.
3. Finally, policy makers and stakeholders from academia, scientific publishing, industry, government and other strategic organisations can exploit an increasing number of evaluation instruments being evaluated and catalogued to understand the landscape of AI evaluation much better.

The facets framework can help these groups recognise gaps and limitations in evaluation instruments of AI performance. In this way, it helps stakeholders move beyond unstructured collections of benchmark results by metric, which are typical of the AI evaluation field. These can be useful for meta-analysis but are still lacking structure and insight about the evaluation instruments themselves.

The analysis of rater disagreement across facet values found it tended to reflect rater uncertainty about what evaluation instruments set out to measure or unresolved issues in AI evaluation. Section 4 observed disagreement between CAPABILITY and PERFORMANCE (**Capability**), between SPECIFIC and CONTAMINATED (**Specificity**), and between UNSTRUCTURED and ABLATABLE (**Adjustability**).

Evaluation instruments rated with the CAPABILITY (**Capability**) value were much more likely to be CONTAMINATED (**Specificity**). This may illustrate a difficulty in interpreting what the evaluation instrument designers intended to measure, particularly when measuring AI wider capabilities. The object of an evaluation instrument tended to be clearer to the raters when it was evaluating narrow task performance.

Rater disagreement may also be a sign of unresolved issues in AI evaluation: going from task-oriented evaluation based on performance to more general evaluation instruments lead to an evaluation instrument becoming CONTAMINATED (**Specificity**). For instance, adding many millions of examples can increase coverage. However, this adds problems of specificity and more difficulty in understanding the role each example plays in the overall score being measured by the evaluation instrument.

The most challenging parts of this proposal were:

- Determining the criteria for the inclusion of evaluation instruments.
- Defining facets that were difficult to understand or liable to be confused with others.
- Finding a protocol of application that is both sufficiently robust and can be used by a limited number of raters with restricted resources.

Finally, the categorisation framework for evaluation instruments presented here should be a living framework rather than set in stone. This would allow facets to be added, changed or removed and updated, and the rubric updated, to reflect the evolving nature of the evaluation of AI systems. However, some stability in names, facet values and facet description is needed to compile results of different rating studies over time. This would permit a large increase from the 36 evaluation instruments evaluated here to the order of hundreds in the future, with a more diverse and numerous pool of raters. Thus, rather than a continually evolving framework, it may be more sensible to review it periodically, following “change requests” from the community. A new, numbered version could be produced, with backward incompatibilities explicitly noted. Hopefully, these facets and the rubric describing them can help track the evolution of AI evaluation in the years to come, and identify the facets where changes are happening or should happen.

References

- AERA, APA, NCME (2014), *Standards for educational and psychological testing et al.*, American Education Research Association, [3]
https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf
 (accessed on 28 August 2023).
- ANAC (2021), *ANAC2021 - 12th Automated Negotiating Agents Competition*, [41]
<http://web.tuat.ac.jp/~katfuji/ANAC2021/> (accessed on 18 October 2023).
- Anderson, P. (2018), *Bring Me A Spoon | Matterport3D Simulator and Room-to-Room (R2R) data for Vision-and-Language Navigation*, <https://bringmeaspoon.org/> (accessed on 18 October 2023). [4]
- Assembly (2018), *Robotic Assembly – Recent Advancements and Opportunities for Challenging R&D | NIST*, <https://www.nist.gov/news-events/events/2018/08/robotic-assembly-recent-advancements-and-opportunities-challenging-rd> (accessed on 18 October 2023). [30]
- Bellemare, M. et al. (2013), “The Arcade Learning Environment: An evaluation platform for general agents”, *Journal of Artificial Intelligence Research*, Vol. 47, [24]
<https://doi.org/10.1613/jair.3912>.
- Bisk, Y. et al. (2020), *PIQA: Reasoning about physical commonsense in natural language*, [15]
<https://doi.org/10.1609/aaai.v34i05.6239>.
- Bommasani, R. et al. (2021), “On the Opportunities and Risks of Foundation Models”, [23]
<https://arxiv.org/abs/2108.07258v3> (accessed on 26 September 2023).
- Brown, T. et al. (2020), “Language Models are Few-Shot Learners”, *Advances in Neural Information Processing Systems*, Vol. 2020-December, [21]
<https://arxiv.org/abs/2005.14165v4>
 (accessed on 26 September 2023).
- Cohn, A. et al. (2022), “A Framework for Categorising AI Evaluation Instruments”, <http://ceur-ws.org/Vol-3169/> (accessed on 26 September 2023). [6]
- Deng, J. et al. (2010), “ImageNet: A large-scale hierarchical image database”, pp. 248-255, [10]
<https://doi.org/10.1109/CVPR.2009.5206848>.
- Devlin, J. et al. (2018), “Bert: Pre-training of deep bidirectional transformers for language understanding”, *arXiv preprint arXiv:1810.04805*. [47]
- Devlin, J. et al. (2018), “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Vol. 1, pp. 4171-4186, <https://arxiv.org/abs/1810.04805v2> (accessed on 26 September 2023). [20]
- Froleyks, N. et al. (2021), “SAT Competition 2020”, *Artificial Intelligence*, Vol. 301, p. 103572, [28]
<https://doi.org/10.1016/J.ARTINT.2021.103572>.

- Gaggl, S. et al. (2020), “Design and results of the Second International Competition on Computational Models of Argumentation”, *Artificial Intelligence*, Vol. 279, p. 103193, <https://doi.org/10.1016/J.ARTINT.2019.103193>. [26]
- Genesereth, M., N. Love and B. Pell (2005), “General Game Playing: Overview of the AAAI Competition”, *AI Magazine*, Vol. 26/2, pp. 62-62, <https://doi.org/10.1609/AIMAG.V26I2.1813>. [32]
- Harman, D. (1992), “Overview of the First Text REtrieval Conference (TREC-1).”, pp. 1-20, <http://trec.nist.gov/pubs/trec1/papers/01.txt> (accessed on 18 October 2023). [45]
- Hendrycks, D. et al. (2021), “Measuring Coding Challenge Competence With APPS”, <https://arxiv.org/abs/2105.09938v3> (accessed on 26 September 2023). [22]
- Hodaň, T. et al. (2018), “BOP: Benchmark for 6D object pose estimation”, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 11214 LNCS, pp. 19-35, https://doi.org/10.1007/978-3-030-01249-6_2. [42]
- Hsu, C. and B. Sandford (2007), “The Delphi Technique: Making Sense of Consensus”, *Practical Assessment, Research, and Evaluation*, Vol. 12, p. 10, <https://doi.org/10.7275/pdz9-th90>. [5]
- Levesque, H., E. Davis and L. Morgenstern (2011), “The Winograd Schema Challenge”, <http://www.aaai.org> (accessed on 26 September 2023). [12]
- Lifelong Robotic Vision (2023), *Lifelong Robotic Vision | IROS2019 Competition*, <https://lifelong-robotic-vision.github.io/> (accessed on 9 October 2023). [9]
- Linares López, C., S. Jiménez Celorrio and Á. García Olaya (2015), “The deterministic part of the seventh International Planning Competition”, *Artificial Intelligence*, Vol. 223, pp. 82-119, <https://doi.org/10.1016/J.ARTINT.2015.01.004>. [36]
- Maas, A. et al. (2011), *Learning Word Vectors for Sentiment Analysis*, <https://aclanthology.org/P11-1015> (accessed on 26 September 2023). [31]
- Marot, A. et al. (2021), “Learning to run a Power Network Challenge: a Retrospective Analysis”, *Proceedings of Machine Learning Research*, Vol. 133, pp. 112-132, <https://arxiv.org/abs/2103.03104v2> (accessed on 26 September 2023). [34]
- Martínez-Plumed, F. et al. (2021), “Research community dynamics behind popular AI benchmarks”, *Nature Machine Intelligence*, Vol. 3/7, pp. 581-589, <https://doi.org/10.1038/s42256-021-00339-6>. [1]
- MineRL (2023), *MineRL: Towards AI in Minecraft*, <https://minerl.io/> (accessed on 26 September 2023). [37]
- Mishra, S. et al. (2022), “NumGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks”, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 3505-3523, <https://doi.org/10.18653/v1/2022.acl-long.246>. [43]
- Ott, S. et al. (2022), “Mapping global dynamics of benchmark creation and saturation in artificial intelligence”, *Nature Communications*, Vol. 13/1, <https://doi.org/10.1038/s41467-022-34591-0>. [2]

- Panayotov, V. et al. (2015), “Librispeech: An ASR corpus based on public domain audio books”, [27]
ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Vol. 2015-August, pp. 5206-5210,
<https://doi.org/10.1109/ICASSP.2015.7178964>.
- PASCAL (2012), *PASCAL VOC 2012 test Benchmark (Semantic Segmentation) | Papers With Code*, <https://paperswithcode.com/sota/semantic-segmentation-on-pascal-voc-2012> [39]
 (accessed on 26 September 2023).
- Perez-Liebana, D. et al. (2019), “The Multi-Agent Reinforcement Learning in Malm\“O (MARL\“O) Competition”, <https://arxiv.org/abs/1901.08129v1> (accessed on 18 October 2023). [11]
- Rajpurkar, P., R. Jia and P. Liang (2018), “Know What You Don’t Know: Unanswerable Questions for SQuAD”, *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, Vol. 2, pp. 784-789, [33]
<https://doi.org/10.18653/v1/p18-2124>.
- Rajpurkar, P. et al. (2016), *SQuAD: 100,000+ Questions for Machine Comprehension of Text*, Association for Computational Linguistics, Stroudsburg, PA, USA, [17]
<https://doi.org/10.18653/v1/D16-1264>.
- Regnier, R. et al. (2021), “Validation de méthodologies d’évaluation de solutions de désherbage autonomes, dans le cadre des projets Challenge ROSE et METRICS.”, *Revue Ouverte d’Intelligence Artificielle*, Vol. 2/1, pp. 11-32, <https://doi.org/10.5802/ROIA.8/>. [38]
- Renz, J. et al. (2019), “AI meets Angry Birds”, *Nature Machine Intelligence* 2019 1:7, Vol. 1/7, [25]
 pp. 328-328, <https://doi.org/10.1038/s42256-019-0072-x>.
- RGMC (2022), *Robotic Grasping and Manipulation Competition @ ICRA 2022*, [46]
https://rpal.cse.usf.edu/rgmc_icra2022/ (accessed on 18 October 2023).
- Robocup (2023), *RoboCup Standard Platform League*, <https://spl.robocup.org/> (accessed on [7]
 26 September 2023).
- Robocup@Home (2023), *RoboCup@Home – Where the best domestic service robots test themselves*, <https://athome.robocup.org/> (accessed on 9 October 2023). [8]
- Sap, M. et al. (2019), “SocialQA: Commonsense Reasoning about Social Interactions”, *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 4463-4473, <https://doi.org/10.18653/v1/d19-1454>. [16]
- Sutcliffe, G. (2016), “The CADE ATP system competition - CASC”, *AI Magazine*, Vol. 37/2, [44]
 pp. 99-101, <https://doi.org/10.1609/AIMAG.V37I2.2620>.
- Vinyals, O. et al. (2017), “StarCraft II: A New Challenge for Reinforcement Learning”, [40]
<https://arxiv.org/abs/1708.04782v1> (accessed on 26 September 2023).
- Wang, A. et al. (2019), “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”, *Advances in Neural Information Processing Systems*, Vol. 32, [14]
<https://arxiv.org/abs/1905.00537v3> (accessed on 26 September 2023).

- Wang, A. et al. (2018), “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”, *EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop*, pp. 353-355, <https://doi.org/10.18653/v1/w18-5446>. [13]
- WN18RR (2023), *WN18RR Benchmark (Link Prediction) | Papers With Code*, <https://paperswithcode.com/sota/link-prediction-on-wn18rr> (accessed on 26 September 2023). [35]
- Yang, Y., W. Yih and C. Meek (2015), “WIKIQA: A challenge dataset for open-domain question answering”, *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pp. 2013-2018, <https://doi.org/10.18653/v1/D15-1237>. [18]
- Zellers, R. et al. (2018), “From Recognition to Cognition: Visual Commonsense Reasoning”, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2019-June, pp. 6713-6724, <https://doi.org/10.1109/CVPR.2019.00688>. [29]
- Zellers, R. et al. (2018), “SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference”, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 93-104, <https://doi.org/10.18653/v1/d18-1009>. [19]

Annex 6.A. Supplementary tables

Annex Table 6.A.1. Overview of Evaluation Instruments

Acronym	Reference	Type	Domain	Aim
WSC	(Levesque, Davis and Morgenstern, 2011 ^[12])	benchmark, competition	Reasoning	Targets evaluating common sense reasoning as a better alternative to the Turing test.
ALE	(Bellemare et al., 2013 ^[24])	benchmark	Video games	Intended to assess general AI using a variety of video games; exact metrics are unclear.
GLUE	(Wang et al., 2018 ^[13])	benchmark	Language	Measures AI performance in English natural language understanding tasks such as single-sentence tasks, similarity, paraphrasing and inference.
SUPERGLUE	(Wang et al., 2019 ^[14])	benchmark	Language	Measures AI performance in English natural language understanding tasks such as single-sentence tasks, similarity, paraphrasing and inference.
IMAGENET	(Deng et al., 2010 ^[10])	competition	Vision	Assesses AI's visual recognition abilities in object recognition, image classification and localisation amid varied conditions.
AIBIRDS	(Renz et al., 2019 ^[25])	competition	Video games	Evaluates an agent's planning ability in large action spaces by making it play Angry Birds.
ICCM	(Gaggl et al., 2020 ^[26])	competition	Reasoning	Compares the performance finding logical solutions in argumentation tasks.
Robocup	(Robocup, 2023 ^[7])	competition	Robotics	Aims to advance multi-robot systems through soccer matches.
Robocup@home	(Robocup@Home, 2023 ^[8])	competition	Robotics	Assesses AI robots in delivering assistive services for future domestic use.
Librispeech-SL12	(Panayotov et al., 2015 ^[27])	dataset	Language	Provides a free English speech corpus for training/testing speech recognition systems.
GVGAI	(Perez-Liebana et al., 2019 ^[11])	competition	Video games	Targets systems that can excel in multiple video games as a step towards artificial general intelligence.
PIQA	(Bisk et al., 2020 ^[15])	benchmark, dataset	Language, Reasoning	Evaluates language-based physical interaction reasoning for both typical and unconventional object uses.
SAT	(Froleyks et al., 2021 ^[28])	competition	Reasoning	Focuses on improving the performance and robustness of SAT solvers.
VCR	(Zellers et al., 2018 ^[29])	dataset	Reasoning, Vision	Identifies human actions and goals from visual cues.
Assembly	(Assembly, 2018 ^[30])	competition	Robotics	Assesses robotic systems' competencies using formal evaluations to guide development and match user needs.
IMDb	(Maas et al., 2011 ^[31])	dataset	Language	Detects text sentiment.
SocialQA	(Sap et al., 2019 ^[16])	benchmark	Socio-emotional	Measures computational models' social and emotional intelligence through multiple-choice questions.
GGP	(Genesereth, Love and Pell, 2005 ^[32])	competition	Video games	Tests AI's ability to play multiple games.
SQUAD2.0	(Rajpurkar, Jia and Liang, 2018 ^[33])	dataset	Language	Evaluates reading comprehension abilities.

Acronym	Reference	Type	Domain	Aim
ZellersWikiQA	(Yang, Yih and Meek, 2015 ^[18])	benchmark	Language	WIKIQA is a dataset for open-domain question-answering.
sW/AG	(Zellers et al., 2018 ^[19])	dataset, benchmark	Language, Reasoning	Measures grounded commonsense inference by answering multiple-choice questions.
L2RPN	(Marot et al., 2021 ^[34])	competition	Reasoning	Tests AI's ability to solve an important real-world problem for the future.
W/AG Lifelong-Robots	(Lifelong Robotic Vision, 2023 ^[9])	competition	Robotics, Vision	Tests AI's ability to solve an important real-world problem for the future.
WN18RR	(WN18RR, 2023 ^[35])	dataset	Reasoning	Measures success in link prediction tasks without inverse relation test leakage.
Planning	(Linares López, Jiménez Celorrio and García Olaya, 2015 ^[36])	competition	Reasoning	Assesses automated planning and scheduling across different problem families.
MineRL	(MineRL, 2023 ^[37])	competition	Video games	Evaluates the performance of reinforcement learning agents in playing Minecraft.
ROSE	(Regnier et al., 2021 ^[38])	competition	Robotics (non-humanoid)	Measures agricultural robotics' market-related aspects for the near future.
PASCAL-VOC	(PASCAL, 2012 ^[39])	dataset, competition	Vision	Evaluates computer vision tasks like object detection and image segmentation.
Starcraft II	(Vinyals et al., 2017 ^[40])	benchmark, dataset	Video games	Assesses agents in playing Starcraft II, also examines perception, memory and attention.
ANAC	(ANAC, 2021 ^[41])	competition	Language, Reasoning	Evaluates multi-issue negotiation strategies.
BOP	(Hodaň et al., 2018 ^[42])	benchmark	Vision	Measures "object pose estimation" in RGB-D images.
NumGlue	(Mishra et al., 2022 ^[43])	benchmark	Reasoning	Evaluates basic arithmetic understanding.
CASC	(Sutcliffe, 2016 ^[44])	competition	Reasoning	Evaluates theorem proving.
TREC	(Harman, 1992 ^[45])	competition	Language	Evaluates information retrieval technology through adaptive yearly competitions.
Bring-MeASpoon	(Anderson, 2018 ^[4])	Benchmark, dataset	Language, Robotics	Tests an agent's ability to navigate to a goal location in an unfamiliar building using natural language instructions.
RGMC	(RGMC, 2022 ^[46])	competition	Robotics	Assesses robotic grasping and manipulation capabilities.

Note: The yellow shaded evaluation instruments are Batches 1 and 2, green shaded items Batch 3 and blue shaded items Batch 4.

Notes

¹ REAL-LIFE does not mean a final or specific product in operation. It can also happen in early stages of research, such as evaluating prototype chatbots in a real social network.

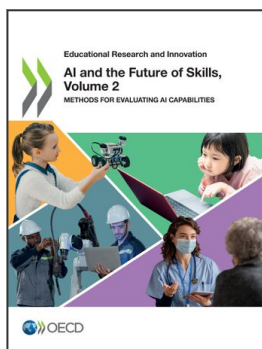
² In AI research, the term “ablatable” refers to a component or feature of a system that can be removed or “ablated” to assess its impact on the system's overall performance.

³ Although PROCEDURAL was coloured, procedural may not always be better and can lead to problems if variations are not in an appropriate proportion. Also, generated data may just lead to a learning algorithm reverse-engineering the generator.

⁴ LOOSE refers to cases when evaluation is open, e.g. a robotic-domain evaluation instrument where a satisfactory interaction with the user is evaluated, but not even a clear questionnaire is defined. FULLY-INDEPENDENT could treat different groups differently if there is a reason for equality of treatment.

⁵ Coverage is about the domain, while Partiality is about how the evaluation instrument may favour some test-takers over others.

⁶ Repositories used were: Papers with code (<http://paperswithcode.com>), Kaggle (<http://kaggle.com>), Zenodo (<https://zenodo.org/record/4647824#.YV7CPdrMKUk>), Electric Frontier Foundation (<https://www.eff.org/ai/>), Wikipedia (https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research), Challenges in Machine Learning (<http://www.chalearn.org>).



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:

Cohn, Anthony G. and José Hernández-Orallo (2023), "A framework for characterising evaluation instruments of AI performance", in OECD, *AI and the Future of Skills, Volume 2: Methods for Evaluating AI Capabilities*, OECD Publishing, Paris.

DOI: <https://doi.org/10.1787/f585066f-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>.