A Framework for Categorising AI Evaluation Instruments

Anthony G Cohn[1], José Hernández-Orallo[2], Julius Sechang Mboli[3], Yael Moros-Daval[2], Zhiliang Xiang[4] and Lexin Zhou[2]

[1]School of Computing, University of Leeds, UK; and the Turing Institute, UK
[2]VRAIN, Universitat Politècnica de València, Spain
[3]Faculty of Engineering and Informatics, University of Bradford, UK
[4]IROHMS, School of Computer Science and Informatics, Cardiff University, UK

Abstract

The current and future capabilities of Artificial Intelligence (AI) are typically assessed with an ever increasing number of benchmarks, competitions, tests and evaluation standards, which are meant to work as AI *evaluation instruments* (EI). These EIs are not only increasing in number, but also in complexity and diversity, making it hard to understand this evaluation landscape in a meaningful way. In this paper we present an approach for categorising EIs using a set of 18 *facets*, accompanied by a rubric to allow anyone to apply the framework to any existing or new EI. We apply the rubric to 23 EIs in different domains through a team of raters, and analyse how consistent the rubric is and how well it works to distinguish between EIs and map the evaluation landscape in AI.

Keywords

Evaluation Instruments, Comparison of Evaluation Instruments, Categorisation of Evaluation Instruments, Artificial Intelligence Evaluation, Future of Skills

1. Introduction

Ever since researchers started building AI systems, they have wanted to evaluate them, either against human benchmarks (such as playing humans experts at Chess or other games) and/or against other AI systems. Finding good benchmarks for evaluating systems, and conducting tests is harder than it might seem, particularly since we believe we have good methods for evaluating human intelligence, via standard tests and examinations.

There have been many tests proposed for evaluating AI systems. Probably the most famous of these of course is known as the *Turing Test*[1]. There have been various Turing Test competitions, of which the best known is the annual Loebner Prize competition; the results have been sometimes entertaining, and a way of promulgating ideas about AI to the general public, but it is hard to argue that any real important progress in AI has been demonstrated by the entrants. In fact, Turing himself never proposed the test as a serious way of measuring AI systems or of measuring progress, as Schieber [2] observes, adding, it is "misguided and inappropriate" ([3, 4]). Instead he argues for new "inducement prize" contests. According to Schieber, these are "award programs established to induce people to solve a problem of importance by directly rewarding the solver". Perhaps the most famous historical examples are the Longitude Rewards offered by the UK government in 1714. A current example is the $5M IBM Watson AI XPRIZE which "challenges teams to demonstrate how humans can work with AI to tackle global challenges". Further discussion on the use of competitions, benchmarks and datasets in evaluating AI systems can be found in [5].

The situation today is that there are thousands of challenges in almost all areas of AI. They are increasing in complexity and diversity, as AI techniques evolve likewise. Because of this, it is hard to analyse this evaluation landscape in a meaningful way. Motivated by this need, we present and discuss an approach to categorising benchmarks, competitions, tests and evaluation standards, jointly referred to as AI *evaluation instruments* (EI). We do this categorisation via a set of 18 *facets*, which we believe will be valuable in distinguishing and evaluating different proposals for evaluating AI systems. These facets, and an accompanying rubric to facilitate choosing appropriate values, are described in section 2.

We will classify EIs using the facets in order to (a) evaluate how well the facets work in general and (b) to what extent they help mapping the landscape of EIs and

IJCAI2022 Workshop on AI Evaluation Beyond Metrics (EBeM'22), July 24, 2022, Vienna, Austria
a.g.cohn@leeds.ac.uk (A. G. Cohn); jorallo@upv.es (J. Hernández-Orallo); mboli4god@gmail.com (J. S. Mboli); ymordav@inf.upv.es (Y. Moros-Daval); xiangz6@cardiff.ac.uk (Z. Xiang); lzhou@inf.upv.es (L. Zhou)
https://eps.leeds.ac.uk/computing/staff/76/professor-anthony-tony-g-cohn-freng-ceng-citp/ (A. G. Cohn); http://josephorallo.webs.upv.es/ (J. Hernández-Orallo); https://jsmboli.github.io/jsmboli/ (J. S. Mboli); https://zl-xiang.github.io/ (Z. Xiang); https://lexzhou.github.io/ (L. Zhou)
0000-0002-7652-8907 (A. G. Cohn); 0000-0001-9746-7632 (J. Hernández-Orallo); 0000-0003-1708-3052 (J. S. Mboli); 0000-0001-5442-2055 (Y. Moros-Daval); 0000-0002-0263-7289 (Z. Xiang); 0000-0003-1161-4270 (L. Zhou)

distinguish their differences. This may help inform how much we can translate from the facet values to guide the design of future EIs. We do not imagine there can be a single universal evaluation instrument, or even a battery for each domain (vision, reasoning, etc.); certainly that ideal has eluded the community so far. We do not even aspire to find facet values that are valid for all EIs but our proposed work may help in directing future efforts in the evaluation of AI systems.

Since it is infeasible in a reasonable amount of time to apply this categorisation to the thousands of EIs in the literature, here we cover 23 EIs (see Table 2). By evaluating a reasonable number of carefully chosen examples, we hope to give a fair picture of the extent to which the aspects of AI appraised by the facets are being tested in the selected examples. Beyond the insights that we extract from this selected set of EIs, this paper and the rubric we have developed for the different facets should serve as a reference for third parties (e.g., other researchers) to analyse other EIs.

The rest of the paper is organised as follows. Section 2 presents the 18 facets and a *rubric* which explains how facet values should be chosen. Next, in section 3, we discuss the criteria for selecting the 23 EIs and the methodology the raters used to apply the rubric. Section 4 discusses the level of disagreement between raters for each facet and EI, and how the methodology and the number of raters was adapted based on these observations. Section 5 analyses the ratings of the 23 EIs, and what they reveal about this group of EIs. Finally, section 6 closes with some general discussion and possible future work.

2. Characterising AI Evaluation Instruments

We looked for existing features or dimensions to characterise EIs, but unfortunately we did not find any systematic account in AI, other than concepts such as reproducibility, realism, coverage and specificity, usually referred to with other names and applied to a single EI. We found more dimensions and a more systematic coverage of evaluation instruments in the area of psychological testing. As a result, we have introduced a new set of facets, but when possible, the terminology is based on the common use in AI, but also incorporating terms and concepts from the Standards for Educational and Psychological Testing by the American Educational Research Association [6].

The following list[1] proposes 18 facets to characterise existing and future EIs for AI. Each facet is followed by the options in brackets. Some options indicate '(specify)', which means that the rater must indicate a (freetext) value for that option. The full description of the facets usually include some examples and further clarifications[2]. Here we only include the basic definition of each of them. We use colours (blue and black) that are indicative, with blue referring to the preferred or most challenging case, *in general*. However, for some facets a blue value may make no sense, or we do not believe that one value is 'better' than any other, so these facets have no coloured facet value(s).

- **Vp - 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?
- **Vc - Capability** [TASK-PERFORMANCE (specify), CAPABILITY (specify)]: Does the EI just measure observed (aggregated) performance on a TASK (e.g., protein folding, credit scoring) or is the EI *designed* to also measure a CAPABILITY (e.g., object permanence, dealing with negation)?
- **Vf - 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?
- **Vo - Coverage** [BIASED (specify), REPRESENTATIVE]: Does the EI cover a BIASED or unbiased (REPRESENTATIVE) distribution of what is meant to be measured?
- **Vs - Specificity** [SPECIFIC, CONTAMINATED]: Are the results precisely aligned with what is meant to be measured or contaminated by other skills or tasks?
- **Vl - Realism** [TOY, GAMIFIED, REALISTIC, REAL-LIFE]: To what extent is the EI a toy problem, a complex gamified problem, is it a realistic setting (e.g., but still in a simulated scenario, a lab or testing facility) or is the evaluation itself happening in real life[3]?
- **Cj - 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?
- **Cc - Containedness** [FULLY-CONTAINED, PARTIAL-INTERFERENCE (specify), NOT-CONTAINED (specify)]: Once started, is the testing isolated from external factors or interference possibly having an effect on the results (human participants, online data, weather, etc.), or is there some partial interference not affecting the results significantly or is it dependent of external resources and conditions?

[1] Each facet has both a name and a two letter acronym, whose initial letter is V, C or F, the reason for which will become clear later.

[2] The latest version of the rubric can be found in https://tinyurl.com/mr2bv5hb

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

- **Cp - Reproducibility** [NON-REPRODUCIBLE, STOCHASTIC, EXACT]: Is the evaluation non-reproducible, with results biased or spoiled if repeated; does the EI have stochastic components leading to different interactions; or are the results completely reproducible, i.e. can exactly the same test (inputs, interaction, etc.) be generated again for another (or the same) competitor?
- **Cl - 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 EI is not clear then N/A may be the most appropriate facet value.
- **Cv - 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 they been created (e.g., using procedural generation[4])?
- **Ca - Adjustability** [UNSTRUCTURED, ABLATABLE, ADAPTIVE]: Is the analysis of results on the set of instances unstructured; or has the EI 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)?
- **Fn - Antecedents** [CREATED, RETROFITTED (specify)]: Is it devised on purpose for AI or adapted from tests designed to test humans.
- **Fm - Ambition** [SHORT, LONG]: When the EI was created, was it aiming at the short term (improving on the SOTA) or long term (more ambitious goals)?
- **Fp - Partiality** [PARTIAL (specify), IMPARTIAL]: Does the EI favour particular technologies, conditions or cultures that should not have an influence on the result of the evaluation[5]?
- **Fo - Objectivity** [LOOSE, CUSTOMISED, FULLY-INDEPENDENT]: Is it loosely defined, customised to each participant or does the EI have a predetermined independent specification[6].
- **Fr - Progression** [STATIC, DEVELOPMENTAL]: Is the score measuring a capability at one particular moment or is it evaluating the development of the capability of the system within the test?
- **Fu - Autonomy** [AUTONOMOUS, COUPLED (specify), COMPONENT]: Is it measuring an autonomous system, coupled with other systems (e.g., humans) or as a component?

The facets above can be grouped into three main categories following the three main groups given by the Standards for Educational and Psychological Testing [6]: validity, reliability/precision and fairness. We use these three major groups to give some structure to the facets above. Roughly, these groups deal with what is measured, how it is measured and who is measured, respectively.

- Validity group (Does it measure what we want to measure?): **Vp**, **Vc**, **Vf**, **Vo**, **Vs**, **Vl**
- Consistency (Reliability/Precision) group (Does it measure it effectively and verifiably?: **Cj**, **Cc**, **Cp**, **Cl**, **Cv**, **Ca**
- Fairness group (Does it treat all test takers equally?): **Fn**, **Fm**, **Fp**, **Fo**, **Fr**, **Fu**

Some of these are closely related, such as {**Cv**,**Ca**,**Vo**} or {**Fo**,**Cp**}. The term *accommodation* in [6] 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". This is related to **Vs**, **Cv**, **Fo** and **Cc**, and also to the term "measurement invariance", which is very important here to see if accommodations of the same test could evaluate the same construct for different AI systems and even humans.

3. EI Selection and Rating Methodology

Now that the facets and the rubric have been explained, we proceed to discuss how the EIs were selected, what the final selection was, and what protocol we followed in assigning EIs to the raters.

3.1. EI Selection

We considered evaluation instruments with the following criteria for inclusion:

- Potential interest to understand the future of AI skills: An EI might be regarded as being of interest if systems which perform well on it can be regarded as indicating a noteworthy change in the capabilities of AI in general. In other words, progress in this EI requires significant enhancement of AI techniques beyond the specific requirements of the EI.
- Diversity in the kind of task: We tried to cover a variety of domains, formats and types of problems (vision, natural language, competitions, datasets, supervised, etc).

[4] Although we have coloured PROCEDURAL, we recognise that 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.

[5] **Vo-Coverage** is about the domain, whilst **Fp-Partiality** is about how the EI may favour some test-takers over others.

[6] LOOSE refers to cases when evaluation is very open, e.g., a robotic-domain EI where we evaluate on a satisfactory interaction with the user, but not even a clear questionnaire is defined. FULLY-INDEPENDENT could treat different groups differently if there is a reason for equality of treatment.

Level	2 options	3 options	4 options	Total
Consistently Agreed	**Fr**, **Fn**	**Vp** , **Cj**, **Cc**, **Fo**, **Fu**	-	7
Moderately Agreed	**Vf**, **Fp**	**Cp**, **Cv**	**Vl**	5
Often Diverged	**Vc**, **Vo**, **Vs**, **Fm**	**Cl**, **Ca**	-	6

Table 1
Level of agreement for the 18 facets, according to the number of options for each facets.

- Popularity: How many teams have already used this EI? How many published papers refer to it? We can use proxies for this, such as citations to the original papers introducing the EI, the number of results on websites such as paperswithcode.com. We also have to consider that industry-related EI may be less popular than research-oriented EIs. However, given the number of EIs selected, we repeat domains and cover just a few areas (e.g., NLP, vision, robotics) without being comprehensive for all possible domains.
- Currency: we prefer EIs still in active use or recently introduced, rather than those which have fallen out of use.

The source of the EIs was mostly repositories[7] and surveys, institutions such as NIST[8] and LNE[9], and competitions at AI conferences. Then, we identified possible gaps in terms of domains or whether we expect that the answers for some facets are going to be too similar. We also considered whether we would expect to get diversity in the values in blue for the facets, so that we get different levels of quality according to this colour code. Note that at the time of selection we could of course only roughly estimate how many blue categories we might get for each EI. Since we expected to learn more about the categorising of EIs as categorisation proceeded, we did not choose all EIs in advance but selected them incrementally. The 23 selected EIs are shown in Table 2.

These EIs cover a good distribution of benchmarks, competitions and datasets, although some of them can be considered to be in two of these categories. The term 'test' to refer to an EI is less usual. About half of the 23 EIs require the use of language in the inputs and/or outputs, and about one half of them require some kind of perception (mostly computer vision), with some overlap in these two groups. Only a few of the EIs are related to navigation and robotics, in virtual (e.g., video games) or physical environments, and a small number are related to more abstract capabilities or problems related to planning or optimisation.

[7] http://paperswithcode.com, http://kaggle.com, https://zenodo.org/record/4647824#.YV7CPdrMKUk, https://www.eff.org/ai/metrics, https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research, http://www.chalearn.org.

[8] https://www.nist.gov/programs-projects/ai-measurement-and-evaluation

[9] https://www.lne.fr/en/testing/

3.2. Rating Methodology

We devised a protocol to refine and validate the rubric, but also to cover as many EIs as possible, according to the number of raters we had available. We explain the protocol below, but we note that this protocol can be adapted to other situations or can incorporate ideas from consensus-based ratings or the Delphi method [30]. First, two of the authors of this paper (A.C. and J.H-O.) acted as coordinators for the rating process. A total of four raters were chosen. Raters were AI-related undergraduate and graduate students, and were recruited through a selection process and interviews. They are the other four authors of this paper (J-S.M., Y. M-D, Z.X. and L.Z.). Once the raters were appointed, each rater was given some meta-information about each EI (acronym, name, major sources, what it measures, etc.) and had to complete some other general information about each EI. They were also asked some information about their own completion, such as time taken (in hours).

We established three batches, covering 2, 11 and 10 EIs respectively, in the order they are presented in Table 2. The first two EIs had already been used by the coordinators in developing the list of facets and their values. All the subsequent raters started off on these two EIs too and were given feedback on their chosen values before proceeding to any further EIs. We refer to these two EIs as "Batch 1". The next 11 EIs are referred to as "Batch 2". These two batches were done by all four raters, independently. After the analysis of consistency we deemed sufficient to only have two raters per EI. Then, a final set of 10 EIs, referred to as "Batch 3", were each rated by just two raters, for reasons of economy, since we already had reasonable inter-rater consistency after the end of batches 1 and 2. The two raters for each EI were assigned so that all raters would have five EIs, and across their five EIs, they co-rated with all the other three raters (i.e., one EI with one other rater and two EIs with each of the other raters). In this first stage, they worked independently, not sharing values for any of the facets, and only reporting questions and partial results to the coordinators.

There were some changes of the rubric between batches, especially clarifying the description of some of the facets, and in a few cases, changes in the number

Table 2
EIs given to raters and included in our analysis.

Acronym	Type	Domain	Aim	Year
WSC [7]	test, benchmark & competition	LU, CS, reasoning	It was specifically targeted to evaluate common sense reasoning, as an alternative to the Turing test, arguing conceptual and practical advantages	2016
ALE [8]	benchmark	VG; navigation; perception	The original goal was to evaluate "general, domain-independent AI technology", by using a diversity of video games, although what it measures more specifically is unclear.	2013
GLUE [9]	benchmark	LU; text retrieval; world knowledge	The goal of GLUE and superGLUE (an improvement/modified version of GLUE) is to measure the performance (e.g. accuracy, F1-score) of an AI system in natural language understanding tasks (Single-Sentence Tasks, Similarity and Paraphrase Tasks, and Inference Tasks) in English.	2018
SUPERGLUE [10]	benchmark	video games; navigation; perception	The goal of GLUE and superGLUE (an improvement/modified version of GLUE) is to measure the performance (e.g. accuracy, F1-score) of an AI system in natural language understanding tasks (Single-Sentence Tasks, Similarity and Paraphrase Tasks, and Inference Tasks) in English.	2019
IMAGENET [11]	competition	image classification; object recognition; object localisation	Aims to measure the visual recognition capability for object recognition, image classification, and object localisation. The images can contain different numbers of objects (e.g. mammal, bird, fish, vehicle, furniture, tool, flower, fruit, etc.), occlusions, and clutters (i.e. diversity and noise).	2010
AIBIRDS [12]	competition	CV, VG, KRRP	Measures the planning capability of an agent in a large action space, without knowing of the physical parameters of objects, situation given by Angry Birds.	2010
ICCMA [13]	competition	reasoning; AA, CL	Aims to measure/compare the performance of different solvers regarding argumentation (particularly, reasoning problem that requires logic).	2015
Robocup SPL[14]	competition	RCRPVMASS	The aim is to measure & promote improvements in multi-robot (humanoid) systems by playing soccer matches with robots	1998
Robocup@home [15]	competition	HRIC, NMDE, CV, ABP.	aims to measure the performance of the developed AI robots in providing service with assistive robot technology with high relevance for future personal domestic applications.	2006
Librispeech-SL12 [16]	dataset	speech recognition	Aims to provide freely available read speech corpus in English that is suitable for training and testing speech recognition systems.	2015
GVGAI [17]	competition	VG;general AI; PN	Aimed to systems that can perform well in multiple video games, possibly without knowing the game in advance and with little to no specific domain knowledge, as an approximation to artificial general intelligence	2014
PIQA [18]	benchmark dataset	PCU, NLP, reasoning	Aims to measure physical interaction reasoning about both the prototypical use of objects (e.g., shoes are used for walking) and non-prototypical but practically plausible use of objects (e.g., shoes can be used as a doorstop). It targets language representations of knowledge traditionally only seen or experienced.	2019
SAT [19]	competition	boolean satisfiability	Aims to keep progress & further improve the performance & robustness of SAT solvers, with a history dating back to the early 90s, thanks to the persistent efforts of the SAT community.	2002
VCR [20]	dataset	CR; cognition; VR	It aims to measure the ability to infer what is happening in a picture (people's actions, goals, etc.) from visual signs which are obvious for humans.	2019
Assembly [21]	competition	RM, ARH, MPLT, DiHM, RGVELO, Anthropomorphic	Identifying key competencies and characteristics of robotic systems using a robust set of formalized evaluations and benchmarks. To help to match robotic hand capabilities to end-user needs as well as to help provide developers and researchers insight for improving their hardware and software designs	2017
IMDb [22]	dataset	NLP	Detecting the sentiment of a piece of text	2011
SocialIQA [23]	benchmark	SI, SIn, EI, IR	Aimed to measure the social and emotional intelligence of computational models through multiple choice question answering	2019
GGP [24]	competition	game playing	General game playing (GGP) is the design of artificial intelligence programs to be able to play more than one game successfully.	2005
SQUAD2.0 [25]	dataset	reading comprehension; NLP	It aims to measure reading comprehension abilities that allows a system to get a correct answer to a given question when the solution can be extracted from the text or abstain from answering otherwise	2018
WikiQA [26]	benchmark dataset	NLP	WIKIQA is a dataset for opendomain question answering	2014
sW/AG [27]	dataset, benchmark	NLI, CR	Aims to evaluate the performance of a system in grounded commonsense inference (reasoning about a situation and anticipate what might come next) by answering multiple choice questions	2018
L2RPN [28]	competition	SG, AI, PG, PN	This challenge aims at testing the potential of AI to address this important real-world problem for our future.	2012
Lifelong-Robots [29]	competition	robotics, CV, RV	Provides a robotic vision dataset collected from real time environments to accelerate both research and applications of visual models for robotics.	2019

Abbreviations: HRIC = Human-Robot-Interaction and Cooperation; NMDE = Navigation and Mapping in dynamic environments; CV = Computer Vision, ABP = Adaptive Behaviors, planning; AA = abstract argumentation; CL = computational logic; VG = video games; KRRP = knowledge representation; reasoning; planning; RCRPVMASS robotics; cooperation; real-time planning; vision; multiagent systems; strategy; LU = Language understanding; CS = common sense; RM = Robotics in Manufacturing; ARH = Adaptive Robot hands, MPLT = Manipulation planning based on learning techniques;DiHM = Dexterous in-hand manipulation; RGVELO = Robust grasping with various everyday life objects; SI = social interaction, SIn = social intelligence, EI = emotional intelligence, IR = inferential reasoning; CR = commonsense reasoning;VR = visual recognition; PN = planning and navigation, SG = Smart Grids, PG = Power Grids, PN = Power networks, PCU = physical commonsense understanding, NLI = natural language inference, RV = Robotic vision

and/or name of the options. Whenever a change was introduced, the raters were informed and had to revisit their ratings for previous batches.

In a second and final stage of the process, the coordinators allowed the raters to exchange opinions, but they were not asked to reach a consensus, just to identify possible misunderstandings. From this discussion, a few ratings were modified. Unless explicitly stated, we refer to these final ratings in the rest of the paper.

4. Analysis of Rater Consistency

As noted above, the 1st and 2nd batches differ from batch 3 because the former had four raters whilst the latter only two. Thus, in the former case, a majority agreement can be formed with three or four raters agreeing, whilst in batch 3 only when both raters agree; hence 'majority' is less statistically significant for the 3rd batch. For simplicity, we will use round A and round B respectively when referring to the first two batches and the 3rd batch. As shown in Figure 1, the level of agreement coincides to a great extent when comparing the results from all batches (Figure 1, top) with the individual ones from round A (Figure 1, middle) and round B (Figure 1, bottom). It can be expected that those facets with more possible values (4) might have more disagreements than those with only two possible values, simply for statistical reasons. We can see that in fact this is not having a big effect, as shown in Table 1.

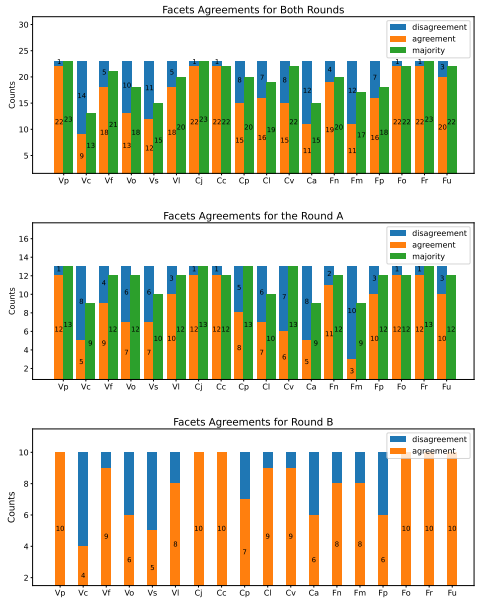

Figure 1: Agreements on facet value ratings for the 23 EIs and rounds A and B.

The pattern of agreement or disagreement amongst the raters tend to vary depending on several factors such as facet complexity, available information on the EI, and so on. In particular, we observe the following:

- **Fr**, **Fn**, **Vp**, **Cj**, **Cc**, **Fo**, **Fu** are *consistently agreed* across all batches, with very few disagreements.
- **Vf**, **Vl**, **Cp**, **Cv**, **Fp** appear to be *moderately agreed* and supported by a majority ($\geq$ 75%). Notably, **Vl** has the largest number of value options, but still agreed well by a majority.
- While selections on **Vo**, **Vs** and **Cl** with binary options, are two of the least agreed ones.

It is not surprising that some of the facets consistently reached consensus considering the facet values tend to distribute towards one single selection (detailed in Section 5). For instance, as we will see in the following section, RESEARCH is picked for the **Vp** facet with only one disagreement for all rounds. This might reflect the fact that some EIs do not have much variability in their options. For example, most EIs are indeed proposed for the purpose of research (**Vp**), and given the low variability in the values there cannot be much disagreement (the variance of a Bernoulli distribution). As the variability of facets increases, choosing answers for the facets might require more EI-specific domain knowledge from the raters. For instance, to make justifiable decisions for facets like **Vo** and **Vs**, raters often need to seek related literature for support when the answers were not clear from the specifications of EIs. Whether an EI is specific (**Vs**) and general (**Vo**) enough for the measuring of certain capabilities is indeed hard to judge depending solely on the specifications. As such, information that is extracted from different sources might lead to disagreements on selections.

Moreover, subjectivity of a facet could also contribute to value divergences. This might be a reasonable explanation for inconsistent selections in **Vc**, **Ca** and **Fm** since they allow raters more space for subjective interpretations. While relevant information w.r.t. **Vc** and **Fm** is often stated in the EI specifications, these statements can somehow be interpreted in different degrees or ways. For example, an EI 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. Having both option variability and subjectivity made the three facets the least agreed ones. Also, some facets are related, and a disagreement in one may be accompanied with disagreement in others. For instance, when TASK-PERFORMANCE is selected for **Vc**, the value of the **Vs** facet is more likely to be SPECIFIC. As such, **Vs** is more likely to be diverged if disagreement occurred on **Vc**. This might also account for the high diverging rate of facets in the Validity group.

In summary, apart from the statistical reason given by the number of values and their variability, the causes for disagreement can be grouped into three blocks:

- Similarity between facet values: The closeness or similarity between facet options might have also reduced the chance of picking the right option. For example, for the facet **VI** - Realism has four options (TOY, GAMIFIED, REALISTIC and REAL-LIFE), and it is not always easy to distinguish between REALISTIC and REAL-LIFE.
- Insufficient Details: For many EIs, the information or details provided by the organisers of the competition, the test or the datasets in the EI is not sufficient to understand what the EI is actually measuring. Other EIs are well documented and have published articles that make it easy to obtain meta-information and the facets values for such EIs.
- Conflicting Information: One of the factors that did not help is the source of information about each EI. For some EIs, there is perhaps too much information and many papers using them, and they do not always understand the same thing or use it in the same way. One paper or website might be talking about task performance while other sources talk of capabilities or both.

Overall, given these sources and level of disagreement, as shown in Figure 1, we considered the rubric sufficiently validated to move from round A to round B with fewer raters, and for the analysis in the next section.

5. Analysis of Results

Herein, we break down the results obtained by the raters to describe what they reveal about the 23 selected EIs (Table 2). Figure 2 shows the frequencies of different options of the 23 EIs for each of the 18 facets. The frequency is calculated differently in the first and the second round. In round A, since we have four raters, each counts for 0.25 unit of frequency (if all chose the same option, it sums up to 1). In round B, we have two raters, each counts for 0.5. In total, we have a maximum frequency of 23 in each option.

Validity group (Does it measure what we want to measure?): Nearly all EIs are designed to foster RESEARCH (**Vp**) and use ABSOLUTE metrics (a preferred option in **Vf**). The number of EIs dedicated to measure performance on a concrete task and EIs aiming to measure a capability is similar (**Vc**), which suggests that the field (at least as represented by these 23 EIs) is undecided on whether to evaluate performance or capabilities. In **Vo**, most EIs were classified as REPRESENTATIVE. However, the percentage of BIASED EIs is still significant (circa 25%), suggesting more efforts may be needed to improve the coverage of current (as well as the ones to come) EIs to mitigate/avoid unrepresentative and unreliable assessments. Surprisingly, only around half of EIs were SPECIFIC (**Vs**), i.e., another half were CONTAMINATED. All the EIs that were designed for TASK-PERFORMANCE are always SPECIFIC (this is suggested in the rubric) but more interestingly, most EIs designed to measure CAPABILITY are CONTAMINATED (i.e., the results do not completely align with what is meant to be measured). More effort is needed to encourage reliable and robust methodologies to evaluate the capability of the AI systems, although we recognise sometimes it is inevitably hard to measure reliably certain capabilities (e.g., common-sense reasoning). With regard to realism (**VI**), REALISTIC EIs account for a predominant proportion (circa 80%), implying considerable focus on measuring systems solving practical problems, but the evaluation is not in an actual real-life scenario; thus most EIs focus on evaluating the systems in simulated scenarios or scenarios which are an abstraction of a real-world setting.

Consistency group (Does it measure it effectively and verifiably?): Nearly all EIs are FULLY-CONTAINED (**Cc**), implying current EIs enjoy high independence from external factors during the assessment) and RELIABLE (**CI**), which are desirable features. Regarding **Cj**, most EIs evaluate the systems with an AUTOMATED scoring instead of MANUAL or MIXED. This phenomenon can be double-edged since automated scoring is generally more objective and faster to calculate but also requires a proper definition for the scoring[10]. For instance, how do we use an automated scoring to evaluate whether a robotic dancer or cook is good or bad? This may be easy for some human experts but quite hard to define using a metric. Things become particularly complicated when measuring a special capability, such as common-sense reasoning. In terms of **Cv**, nearly all EIs are FIXED datasets. Almost none had altered the instances by adding post-processing variations or created new to cover a range of variations intrinsically, possibly because using fixed datasets is easier than modifying instances systematically. However, this could obstruct the diversity in the evaluation methodology (e.g., sometimes it would be interesting to see how the system's performance varies by adding noise to the data to test the model's robustness). Surprisingly, most EIs are UNSTRUCTURED or ABLATABLE (**Ca**), but almost none are ADAPTIVE. This might be because adaptive tests are much more difficult to operate and require an understanding of what the most informative instances are.

Fairness group (Does it treat all test takers equally?): EIs that are IMPARTIAL account for 80% of the data (**Fp**), which seems a good indicator. However, the actual value might be even lower since it is often hard to detect impartiality. For instance, in an EI for benchmarking clinical decision support systems, the training

[10] Easy scoring gives an impression of higher objectivity but some subjectivity still exists in the choice of the metric itself. Automated scoring usually helps with repeatability and traceability.

set may only include Latin American patients but there are patients from other regions in the test set. Interestingly, virtually all the analysed EIs are classified as FULLY-INDEPENDENT (**Fo**), as values CUSTOMISED and LOOSE are only 0.25 (i.e., these options were only chosen once). The fact that current EIs have the same predetermined specification for all assessed systems is positive and a characteristic that favours fairness in evaluation. Nearly all EIs evaluate the AI systems statically rather than developmentally, possibly because for many applications we care more about the final performance rather than how the system's performance evolves. Also, it is easier to evaluate the former than the latter. However, DEVELOPMENTAL EIs could give more insights about how the models are learning with variations of the input features and different curricula, detect when and why the things go wrong during the training phase, and the trade-off between number of instances, time and performance.

In summary, in the validity of the EIs, we found that most of the selected EIs that measure a capability do not necessarily measure the capability reliably. Still, these failures could serve as excellent future references for developing more robust frameworks for evaluating capabilities, and more efforts are required in the years to come. Also, we still need to improve the coverage (i.e., representativeness) in the current EIs. In addition to that, the development of more EIs with real-life settings, may encourage the development of AI systems better able to operate in real-life situations.

Regarding the consistency group: albeit most of the selected EIs measure effectively and verifiably, as they are FULLY-CONTAINED and RELIABLE, there is still an evident lack of diversity in the evaluation process. For instance, we may need more EIs focusing on altering instances by adding post-processing variations or creating instances to cover a range of variations intrinsically. Also, more adaptive ways to test a system should be encouraged, in order to evaluate how the system copes in circumstances with different difficulties. Finally, in terms of fairness, the selected EIs enjoy low partiality and high objectivity. However, more efforts are needed in spurring EIs to also focus on evaluating how a system performs during the development process. Furthermore, the community may need more benchmarks that focus on humans and machines working together, since only one out of 23 EIs were done this way.

When looking at the distribution of facet values per EI, we can see that those related to robotics and the physical world (Robocup SPL, Robocup@Home and lifelong-robots) have more variability in **judgeability** (MANUAL becomes more frequent), **realism** (REALISTIC and REAL-LIFE also become more frequent) and **containedness** (PARTIAL-INTERFERENCE becoming more common), as well as **autonomy**, with the COUPLED value being chosen in some of them. One of the most popular EIs in

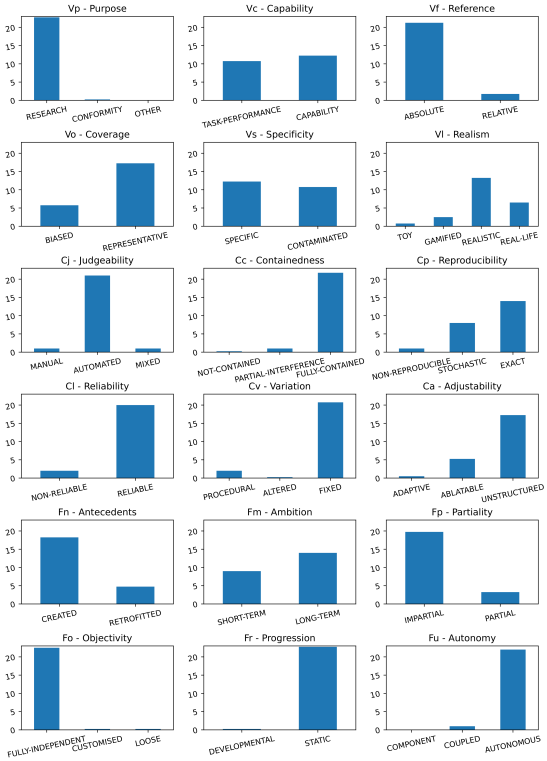

Figure 2: The distribution of the options in all facets.

the history of AI, ImageNet, is the only one where the value PARTIAL is chosen by (at least) half of the raters, and also the one with all BIASED values chosen 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 EI and not identified by everyone immediately. GVGAI is peculiar as a well-thought-out EI, where video games are ablatable by several characteristics or difficulty of the game. This is also going in the direction of being procedural, but still to a limited extent as per the values assigned by the raters for this EI. Finally, those EIs related to natural language, and especially WSC, GLUE, SUPERGLUE, Physical IQa, SocialQA, SQUAD2.0, WikiQA and sW/AG have high degrees of CONTAMINATED values in facet SPECIFICITY. This might be a reflection of how difficult it is to isolate particular capabilities when using natural language, as some basic natural language competency requires many other things. And this is reflected by the success of language models recently doing a variety of tasks [31, 32, 33, 34], since mastering natural language seems to be contaminated by so many other capabilities and skills.

6. Discussion and Conclusions

In section 4 we have seen disagreement between CAPABILITY and PERFORMANCE (**Vc**), between SPECIFIC and CONTAMINATED (**Vs**), and between UNSTRUCTURED and ABLATABLE (**Ca**). The distributions of these facets in section 5 may illustrate a difficulty in interpreting what the EI designers intended, i.e., a lack of clarity in the specification of the EI. It may also be a sign of unresolved issues in AI evaluation: going from task-oriented evaluation based on performance to more general EIs leads to **SPECIFICITY** problems. For instance, adding many millions of examples can help to coverage but comes with problems of specificity and more difficulty in understanding the role each example plays in the overall score being measured by the EI.

Being aware of the consistency issues of the rating methodology, we think the set of facets and associated rubric, as well as the results of the study of 23 EIs reported in this paper, can be useful for three different kinds of users in slightly different ways. First, EI creators can see what design choices in their EI to modify from a first evaluation of its facets and see how it compares to other EIs. For AI system developers, they can choose the right EIs according to the facet values, and better understand what they can expect from the evaluation and what it means exactly. Finally, for policy-makers and stakeholders from academia, scientific publishing, industry, government and other strategic organisations, an increasing number of EIs being evaluated and catalogued can serve to understand the landscape of AI evaluation much better. This can help them recognise gaps and limitations, beyond the unstructured collections of benchmark results by metric that have become very useful for meta-analysis but still lacking structure and insight about the EIs themselves.

In fact, there have been several studies focusing on numeric comparison and the evolution of performance for a range of EIs [35, 36]. These studies see the evolution of the progress of AI systems according to some metrics, but we need more analysis on how the evaluation instruments (benchmarks, competitions, standards, tests, etc.) are also evolving, and whether they are meeting the demands of a more comprehensive evaluation beyond some simple metrics. This was our main motivation.

We have faced some difficulties in determining the criteria for inclusion of EIs, the isolation of some facets that were difficult to understand or confused with others, and finding a protocol of application that is sufficiently robust but at the same time requiring a limited number of raters and other resources. We plan this setting to be a live endeavour, with some facets being added, changed or removed in new versions of the rubric. However, some stability in names, facet values and facet description is needed to be able to compile the results of different rating studies over time, increasing from the 23 EIs evaluated here to the order of hundreds in the future, with a more diverse and numerous pool of raters. As an immediate continuation of this work ourselves, we plan to apply the rubric to further EIs. We hope 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.

Acknowledgments

We thank the anonymous reviewers for their comments. The development of this rubric was performed in the context of the OECD AI and Future of Skills project. Several versions of the facets were discussed in a series of meetings within the project, and especially two meeting in July 5th 2021 and October 26th, where we presented preliminary versions of this rubric. In particular, we thank the OECD team (Stuart Elliott, Abel Baret, Margarita Kalamova, Nóra Révai, Mila Staneva) and the rest of experts and participants (Guillaume Avrin, Lucy Cheke, Kenneth D. Forbus, Yvette Graham, Patrick Kyllonen, Elena Messina, Britta Rüschoff, Michael Schönstein, Jim Spohrer and Swen Ribeiro). We also thank the OECD for the funding which made this work possible as well as their encouragement.

References

[1] A. Turing, Computing machinery and intelligence, Mind 59 (1950) 433.

[2] S. M. Shieber, Principles for designing an AI competition, or why the Turing Test fails as an inducement prize, AI Magazine 37 (2016) 91–96.

[3] S. M. Shieber, Lessons from a restricted Turing Test, Commun. ACM 37 (1994) 70–78.

[4] P. Hayes, K. Ford, Turing test considered harmful, in: International Joint Conference on Artificial Intelligence (IJCAI), 1995, pp. 972–977.

[5] A. G. Cohn, On evaluating artificial intelligence systems: Competitions and benchmarks, in: AI and the Future of Skills, Volume 1 Capabilities and Assessments, OECD, 2021, pp. 238–251.

[6] AERA, APA, NCME, et al., Standards for educational and psychological testing, American Educational Research Association, 2014.

[7] H. J. Levesque, The Winograd Schema Challenge, in: Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011, AAAI, 2011. URL: http://www.aaai.org/ocs/index.php/SSS/SSS11/paper/view/2502.

[8] M. G. Bellemare, Y. Naddaf, J. Veness, M. Bowling, The arcade learning environment: An evaluation

platform for general agents, J. Artif. Intell. Res. 47 (2013) 253–279. URL: https://doi.org/10.1613/jair.3912. doi:10.1613/jair.3912.

[9] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019. URL: https://openreview.net/forum?id=rJ4km2R5t7.

[10] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, Superglue: A stickier benchmark for general-purpose language understanding systems, in: H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 3261–3275. URL: https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

[11] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, IEEE Computer Society, 2009, pp. 248–255. URL: https://doi.org/10.1109/CVPR.2009.5206848. doi:10.1109/CVPR.2009.5206848.

[12] J. Renz, X. Ge, M. Stephenson, P. Zhang, AI meets angry birds, Nat. Mach. Intell. 1 (2019) 328. URL: https://doi.org/10.1038/s42256-019-0072-x. doi:10.1038/s42256-019-0072-x.

[13] S. A. Gaggl, T. Linsbichler, M. Maratea, S. Woltran, Design and results of the second international competition on computational models of argumentation, Artif. Intell. 279 (2020). URL: https://doi.org/10.1016/j.artint.2019.103193. doi:10.1016/j.artint.2019.103193.

[14] The robocup standard platform league, https://spl.robocup.org/, 1998.

[15] The robocup@home league, https://athome.robocup.org/, 2006.

[16] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An ASR corpus based on public domain audio books, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015, IEEE, 2015, pp. 5206–5210. URL: https://doi.org/10.1109/ICASSP.2015.7178964. doi:10.1109/ICASSP.2015.7178964.

[17] D. Perez-Liebana, S. M. Lucas, R. D. Gaina, J. Togelius, A. Khalifa, J. Liu, General video game artificial intelligence, Synthesis Lectures on Games and Computational Intelligence 3 (2019) 1–191. https://gaigresearch.github.io/gvgaibook/.

[18] Y. Bisk, R. Zellers, R. LeBras, J. Gao, Y. Choi, PIQA: reasoning about physical commonsense in natural language, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 7432–7439. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6239.

[19] N. Froleyks, M. Heule, M. Iser, M. Järvisalo, M. Suda, SAT competition 2020, Artif. Intell. 301 (2021) 103572. URL: https://doi.org/10.1016/j.artint.2021.103572. doi:10.1016/j.artint.2021.103572.

[20] R. Zellers, Y. Bisk, A. Farhadi, Y. Choi, From recognition to cognition: Visual commonsense reasoning, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 6720–6731. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Zellers_From_Recognition_to_Cognition_Visual_Commonsense_Reasoning_CVPR_2019_paper.html. doi:10.1109/CVPR.2019.00688.

[21] Assembly performance metrics and test methods, https://www.nist.gov/el/intelligent-systems-division-73500/robotic-grasping-and-manipulation-assembly/assembly, 2018.

[22] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA, The Association for Computer Linguistics, 2011, pp. 142–150. URL: https://aclanthology.org/P11-1015/.

[23] M. Sap, H. Rashkin, D. Chen, R. LeBras, Y. Choi, Socialiqa: Commonsense reasoning about social interactions, CoRR abs/1904.09728 (2019). URL: http://arxiv.org/abs/1904.09728. arXiv:1904.09728.

[24] M. R. Genesereth, N. Love, B. Pell, General game playing: Overview of the AAAI competition, AI Mag. 26 (2005) 62–72. URL: https://doi.org/10.1609/aimag.v26i2.1813. doi:10.1609/aimag.v26i2.1813.

[25] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100, 000+ questions for machine comprehension of text, in: J. Su, X. Carreras, K. Duh (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP

2016, Austin, Texas, USA, November 1-4, 2016, The Association for Computational Linguistics, 2016, pp. 2383–2392. URL: https://doi.org/10.18653/v1/d16-1264. doi:10.18653/v1/d16-1264.

[26] Y. Yang, W. Yih, C. Meek, WikiQA: A challenge dataset for open-domain question answering, in: L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, Y. Marton (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, The Association for Computational Linguistics, 2015, pp. 2013–2018. URL: https://doi.org/10.18653/v1/d15-1237. doi:10.18653/v1/d15-1237.

[27] R. Zellers, Y. Bisk, R. Schwartz, Y. Choi, SWAG: A large-scale adversarial dataset for grounded commonsense inference, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, 2018, pp. 93–104. URL: https://doi.org/10.18653/v1/d18-1009. doi:10.18653/v1/d18-1009.

[28] A. Marot, B. Donnot, G. Dulac-Arnold, A. Kelly, A. O'Sullivan, J. Viebahn, M. Awad, I. Guyon, P. Panciatici, C. Romero, Learning to run a power network challenge: a retrospective analysis, in: H. J. Escalante, K. Hofmann (Eds.), NeurIPS 2020 Competition and Demonstration Track, 6-12 December 2020, Virtual Event / Vancouver, BC, Canada, volume 133 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 112–132. URL: http://proceedings.mlr.press/v133/marot21a.html.

[29] L. Yang, Sdkd: Saliency detection with knowledge distillation, https://lifelong-robotic-vision.github.io/competition/papers/PekingU_linyang.pdf, 2019.

[30] C.-C. Hsu, B. A. Sandford, The Delphi technique: making sense of consensus, Practical assessment, research, and evaluation 12 (2007) 10.

[31] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[32] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 1877–1901.

[33] D. Hendrycks, S. Basart, S. Kadavath, M. Mazeika, A. Arora, E. Guo, C. Burns, S. Puranik, H. He, D. Song, J. Steinhardt, Measuring coding challenge competence with APPS, 2021. arXiv:2105.09938.

[34] R. Bommasani, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258, 2021.

[35] F. Martinez-Plumed, P. Barredo, S. O. Heigeartaigh, J. Hernandez-Orallo, Research community dynamics behind popular AI benchmarks, Nature Machine Intelligence 3 (2021) 581–589.

[36] A. Barbosa-Silva, S. Ott, K. Blagec, J. Brauner, M. Samwald, Mapping global dynamics of benchmark creation and saturation in artificial intelligence, arXiv preprint arXiv:2203.04592 (2022).