AI Evaluation: On Broken Yardsticks and Measurement Scales

José Hernández-Orallo
Universitat Politècnica de València, Spain
Leverhulme Centre for the Future of Intelligence, Cambridge, UK
jorallo@upv.es

Abstract

AI measurement suffers from a moving-target phenomenon, trying to catch up with accelerating AI research. This is partly caused by the "AI effect" (whenever something is automated, intelligence is no longer considered necessary for it), the "superhuman abyss" (once an AI system reaches superhuman performance for a task, the human yardstick does not extend further), the "resource neglect" (breakthroughs are achieved and celebrated independently of the resources involved), the "specialisation drift" (specific solutions for particular benchmarks are favoured by performance-focused competitions) and the "cognitive-judge problem" (failure to distinguish the cognitive effort that is necessary for producing and verifying instances, frequently relying on human labellers or crowdsourcing). The situation is usually associated with a 'challenge-solve-and-replace' dynamics of AI benchmarks: measurement yardsticks are broken once their upper edge is reached, and replaced by a new benchmark. A deeper understanding of these issues reveals lack of measurement invariance, ill-defined scales, measures without unit, and other critical problems we should identify and address, looking at measurement theory and other disciplines. In this paper we examine the moving-target phenomenon and the causes mentioned above, and several research directions towards better AI measurement, focusing on dimensions, scales and units.

Introduction

AI has already achieved superhuman performance in a wide range of tasks: from particular board games to time series prediction. For example, HRA has achieved the maximum score of 999,990 points for Pac-Man, whereas an average human player scores around 15,693 points, and the best score ever achieved by a human is 266,330 points (Van Seijen et al. 2017). But are any of these scores meaningful extrapolations and useful for a quantitative comparison?

In machine vision, in contrast, images are usually labelled by humans, sometimes through crowdsourcing efforts. As a result, ground truth seems tied to expert, consensus or collective human performance. Better-than-human performance can only happen because the average human makes mistakes. Ultimately, in case of doubt, one or more expert humans would label the image, taking this as ground truth. Can we then define superhuman machine vision?

A first option relies on *modifying* some existing images. For instance, if we add noise to distort the image of a cloud, the ground truth is still a cloud, regardless of what humans see in it. One traditional way of doing this is through psychophysical transformations (rotation, contrast, size, etc.), which are assumed to be more independent of how animals and humans perceive (Rajalingham et al. 2018; Leibo and others 2018). A second option relies on *creating* new images (in real or virtual worlds) from scratch, by varying the number of objects, the similarity between them, the locations, etc. These variations can configure different dimensions of *difficulty*, which ultimately would correspond to the *the dimensions of object recognition*.

Hard as it may seem, we can picture how machine vision, as a problem, can be defined in non-anthropocentric terms. However, some other tasks seem inherently human, such as those dealing with natural language; relying on collected or crowdsourced data and corpora seems inescapable. For instance, in machine translation, it is hard to imagine how humans should not be the *yardstick*, at least in terms of quality. Who –but a human– will judge a translation of Shakespeare's *Measure for Measure* into Mandarin? Not only is the task *produced* by a human –Shakespeare–, but also it has to be *evaluated* by a human, a pundit in Shakespeare, English and Mandarin. Can we think of more challenging translation problems, produced and judged automatically? If this were possible, would we then be able to talk meaningfully about the dimensions of progress beyond superhuman performance?

In the previous domains and some other areas of artificial intelligence, we see this human dependency, in order to produce and verify the test instances that are included in

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

many AI benchmarks. But when human level is reached for a particular benchmark, the yardstick does not work well beyond human level, and new –supposedly more challenging– benchmarks replace the old ones.

In what follows, we explore some causes of this 'challenge-solve-and-replace' phenomenon (Schlangen 2019). We identify at least five causes for this:

- The "AI effect" (McCorduck 2004): whenever something is automated, intelligence is no longer considered necessary for it, and interest usually diverts towards more challenging and still intriguing problems.
- The "superhuman abyss": by this we refer to the observation that, once AI reaches superhuman level for a given task, this human yardstick reaches an abyss in which it does not extend further, or could be extended in many arbitrary and unjustified directions, showing the anthropocentrism of the benchmark (Hernández-Orallo 2017a).
- The "resource neglect": by this we mean that many results in AI that are hailed as breakthroughs are obtained with huge resources in terms of data, compute, supervision and many others (Martínez-Plumed et al. 2018a). Even with Moore's law, many of these breakthroughs cannot be shared generally or converted into commercial products easily.
- The "specialisation drift": by this we mean the conscious or unconscious tendency of AI researchers to specialise to a particular task, or even worse, to overfit to a benchmark (known in other areas as Goodhart's law). This especially applies to machine learning, which has struggled with overfitting and generalisation issues for decades (Sutton 1996; Neyshabur et al. 2017).
- The "cognitive-judge problem": by this we refer to a failure to distinguish the manual or automatic cognitive effort that is necessary for producing and verifying instances. In some domains, more challenging instances require more cognitive effort to be produced and verified, and automated metrics usually fail to do a proper job (Kynkäänniemi et al. 2019), with human judges being necessary in the end.

These causes are also deeply intertwined with the way many benchmarks are built today, as huge collections of instances that are produced by humans, labelled by humans, either inadvertently (e.g., translation) or through more or less explicit crowdsourcing procedures.

In the rest of this paper, we will first explore current benchmarks that are linked to an anthropocentric performance yardstick that gets broken whenever AI gets close to human performance. Then we will explore ways in which these benchmarks could be extended. We will illustrate these extrapolations and the dimensions of many domains in AI, and how each dimension can be based on distorsions or cognitive scales that allow for automated production and verification, even if AI becomes involved in the generation of benchmarks of the future. Finally, we summarise and identify a series of areas where AI evaluation can get inspiration and techniques in the near future, in order to derive dimensions, scales and ultimately units of measurement.

Breaking Yardsticks

Remember the Pac-Man example. Since we know the scores for HRA and humans, we can calculate its 'Absolute Turing Ratio' (Masum, Christensen, and Oppacher 2002), the quotient between the performance of the AI system and humans. Using best human performance, we would get a ratio of approximately 4. Does this mean that HRA is four times better than humans? Of course it does not, neither does it give a clue of how difficult it was to achieve this score. Score scales in games are very arbitrary, in the end. This meaningless scale is also visible in extrapolations about the progress of AI (Shoham 2017), found in numerous indices and repositories (aiindex.org, paperswithcode.com). For instance, Fig. 1 shows the evolution of results for the popular image recognition benchmark ImageNet. The estimation of human performance is shown as a dashed horizontal line (as calculated in (Russakovsky et al. 2015)).

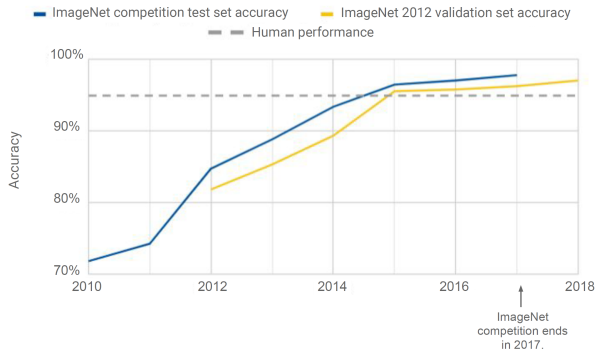

Figure 1: Evolution of the best AI technique on the ImageNet 2012 corpus [image from 2018 AI Index (Shoham et al. 2018)].

We see some extrapolation issues in this plot. If a system reaches 100% correct, this does not mean that it is perfect for image recognition. Indeed, progress becomes less meaningful above average human score (95%). Once this level is met, competitions are soon discouraged and usually discontinued. The proxy has served the community for a while but, very shortly, a more challenging benchmark replaces the old one. This 'challenge-solve-and-replace' evaluation dynamics (Schlangen 2019), or a 'dataset-solve-and-patch' adversarial benchmark coevolution (Zellers et al. 2019) have been reinforced by concepts such as AI-completeness and Human-Level Machine Intelligence (HLMI). AI-completeness (Mueller 1987) was inspired by computational complexity, referring to the class of tasks that require 'human-level intelligence' to be solved. Supposedly, once one of these problems were solved, all other AI-complete problems would be solved.

I will deviate from this view completely. No test in the literature (including the Turing test) has been shown to be sufficient and necessary for certifying the so-called HLMI. Whenever a high score is achieved for some of these purported human-complete tasks (e.g., fooling a number of judges for some time), the reactions vary from criticising

the test implementation to taking the task out of the AI-completeness class (Hernández-Orallo 2000; Vardi 2015). Again, this fits the 'AI effect' narrative well. Even if we jettison the idea of AI-completeness altogether, and refine HLMI as "capable of matching humans in every (or nearly every) sphere of intellectual activity" (Shanahan 2015), we end up with questions such as whether this refers to the average or the best human for each activity, to the point of considering HLMI an "ill-posed" concept (McDermott 2007; Hernández-Orallo 2017a; Martínez-Plumed et al. 2018b).

All things considered, using human intelligence as a yardstick limits our vision of what AI should be, how to devise benchmarks and how to extrapolate beyond them. But how can we compare AI systems and extrapolate their achievements without these human yardsticks?

Exploring Extrapolations

We can classify the tasks we have mentioned above (the Turing Test, Pac Man and ImageNet) into three different categories, as shown in Fig. 2. The first category, 'Ceiling', cannot be extrapolated, either because the ground truth is human or the task measures *humanity*. Apart from the Turing Test, some generators (e.g., realistic human voice generators) fall under this category. The second category, 'Projectional', represents many domains for which, once AI reaches human performance, the score is simply projected, as if the magnitude had a clear meaning. Video game scores are an example of this category. The third category, 'Transitional', adds instance variations humans can conceive and solve to build new dimensions of extrapolation. For instance, we can add Gaussian noise and blur to ImageNet (Dodge and Karam 2017).

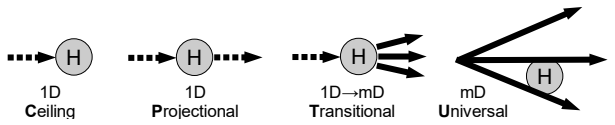

Figure 2: Four different extrapolation categories. The 'Ceiling' (C) category sets humans (H) as a goal and cannot go beyond (e.g., Turing Test). The 'Projectional' (P) aims at humans and then extrapolates the original dimension (e.g., Pac Man). The 'Transitional' (T) extends the space once human performance has been reached (e.g., ImageNet 2012). The 'Universal' (U) defines a (multidimensional) space from the very conception of the task (e.g., brain cancer diagnosis).

The fourth paradigm in Fig. 2 is less usual, but appears in some well-defined scenarios. For instance, the ground truth in brain cancer diagnosis is whether a patient develops cancer in a given time window (e.g., 5 years), independently of what human experts predicted. The instances of this problem are characterised by traits such as kind of cancer, patient group (age, gender, etc.), type of scan, etc. From these traits, we can identify what regions of values make the problem harder, and derive a new group of dimensions that define the multidimensional space of performance. Evaluations should place AI systems –and any particular physician– in this space.

The possibility of extrapolation in several dimensions comes before the question of magnitude. The scales of measurement and units have been brought to AI evaluation more recently (Hernández-Orallo 2017a), but have not always addressed the extrapolation problem, or whether there should be limits or not for some capabilities (Flach 2019b; Hernández-Orallo 2019). We will revisit the question of measurement scales and units later on in this paper.

Of course, many tasks in AI today lack a formulation or the dimensions to fit under the fourth category of 'universal' extrapolation. This is especially the case if we think of dimensions leading to instances that are *cognitively* harder than those humans could solve. In human evaluation, this phenomenon is sometimes referred to as the 'ceiling effect', because it is hard for humans to think of hard instances for humans.

Cognitive Effort: Instance Production, Solving and Verification

For the sake of reaching some target performance level (usually human level) for a famous and challenging benchmark, anything goes. While data (George et al. 2017; Tucker, Anderljung, and Dafoe 2020) and compute (Amodei and Hernandez 2018) have been recognised as two major resources that may explain some recent breakthroughs in AI, there are some other resources that have been usually neglected, such as knowledge, software, hardware, (human) manipulation, network, (physical) time, load and energy (Martínez-Plumed et al. 2018a). All these resources represent an AI footprint that is usually ignored when accounting for the progress (and autonomy) of an AI system. Many of these resources are related. For instance, data collection captures human work, and when data is labelled, this can be considered a variant of human manipulation or supervision. A fair evaluation of how a system (or an AI team) solves a task must account for all these resources, and not only the displayed performance[1].

But even if we consider all the resources needed to *solve* a task, there are some other efforts that still remain eclipsed. One important resource facet to be considered is the effort of *producing* instances. In the cancer diagnosis domain, we know the cost of obtaining and labelling patient data, but in other areas this cost may be associated with producing, e.g., a solvable planning problem, an interesting theorem to prove, a challenging chess position, etc. In other cases, humans are required for *verifying* that the AI system has actually solved a problem or it has a sufficient quality. This is quite common in generative problems, such as realistic image or video generation.

In the end, there might be a combination of human and machine work, and also a combination of automated and manual metrics. However, the meaningful distinction in the future will be how much *cognitive* work is needed to produce and verify the instances. For example, producing and

[1]Some initiatives are putting the focus on this, at least on computation time, such as MLPerf (https://mlperf.org/about/), which replaces and extends DAWNBench (https://dawn.cs.stanford.edu/benchmark/).

Domain	Representative benchmark	Mother distribution (p_M) (application dependent)	Test distribution (p_T) (also used for training)	Instance features	Production	Verification	Proposed dimensions (difficulty metrics)	MT	Ⓗ
Translation	NIST OpenMT (Han 2016)	Texts in human languages and translation queries	A few collected corpora	Length, language, syntactic features, vocabulary, ...	Choose sentence & target language	Human trnsltn. (subj. or scores)	Language divergence, lexical ambiguity, ...	⊖⊖	C
Diagnosis	3064 brain tumor dataset (Cheng et al. 2015)	Human population	Medical samples	Population groups, type of cancer, kind of scan, ...	Test patients and collect	Retrieve class (e.g., after 5 yrs.)	Scan quality, size of spot, antecedent info., ...	?⊕	U
Vehicle driving	K-City (Joerger et al. 2019)	Car trips in the world	Trials in a testbed or restricted area.	Traffic, time, weather, region, type of car, ...	Choose route or destination	Car reaches destination safely	Visibility, traffic density, road state, ...	⊖⊕	T
Face Recognition	DiF dataset (Merler et al. 2019)	Human population	Extracted faces from Flickr sample (YFCC-100M)	Race, age, craniofacial areas, ratios, symmetries, ...	Make photo, add ID and collect	Retrieve ID and check	Trait unspecificity, photo quality, pose, rotation, ...	?⊕	T
Image Generation	CIFAR / ImageNet (Barratt and Sharma 2018)	Meaningful or useful objects in the world	Several image collections	Kind of object, pose, size, location, ...	Choose model, label or traits	Humans or scores (FID, ...)	Texture & colour variation, compositional depth, ...	⊖⊕	C
Board games	AlphaGo/Zero matches (Silver and others 2016)	All human Go players	Some human and machine go players	Elo-like ranking, positions, playing styles, ..	Choose opponent	Opponent plays	Opponent ranking, number of empty cells	⊕⊕	P
Multi-agent pathfinding	Grid-based MAPF (Stern et al. 2019)	Warehouses, cities, etc.	Some grids from games, cities, mazes, ...	Obstacles, topology, agents, etc.	Real cases or generators	Calculate optimality	Bottlenecks, number of agents, ...	?⊕	U
Arcade games	GVGAI (Perez-Liebana et al. 2016)	All arcade games as much as they are played	Selection for GVGAI competition	Number of elements, obstancles, size, ...	Human designer with VGDL	Play game	Reward noise and sparsity, policy complexity, trials, ...	?⊕	P
Language understanding	SuperGLUE (Wang et al. 2019)	Texts & questns in natural language in the world	Collection of texts and questions	Length, language, type of question, , ...	Choose text and human questions	Compare answer	Syntactic and semantic complexity, distractors, ...	⊖⊕	C
Turing test	Loebner's prize (Vardi 2015)	Humans	Chosen humans	Personality, gender, knowledge, capabilities, ...	Humans chat	Humans (peers and judges)	Human capabilities, unpredictability, ...	⊙⊙	C
Language generation	PTB, Wikitext, ... (Radford et al. 2019)	Texts in natural language in the world	A few collected corpora	Topic, style, language, vocabulary, ...	Choose topic, traits or lead text	Humans or perplexity	Semantic depth, style specificity, ...	⊖⊖	C

Table 1: A selection of domains, benchmarks, instance distributions (p_M and p_T), features, how the production and verification is done, proposed dimensions, super/par/subhuman (⊕/⊙/⊖) AI level for p_M and p_T, and tacit extrapolation category (Fig. 2).

verifying translation instances beyond human performance (e.g., translations that are specific for a particular user) may require some advanced cognitive effort.

This is a symptom of a more general phenomenon. Many AI tasks are there simply because some cognitive processes provided their components: summarising a *text*, assigning a *name* to a face, tagging a *category* to an image, fighting a *NPC* in a game, etc. The text, the name, the category, the NPC, etc., are results of human cognitive work. However, as AI exceeds some cognitive capabilities, new objects and tasks will be created by AI too (Carter and Nielsen 2017). The AI tasks of the future will come from complex and costly production processes, taking humans out of the loop.

Let us scrutinise this phenomenon with Table 1, a representative (but non-comprehensive) list of 'domains' and 'benchmarks'. By p_M we denote the distribution of instances $\mu \in M$ that compose a task, the *mother* distribution. For example, for the translation domain, an instance μ is a tuple $\langle S, L_1, L_2 \rangle$, composed of a sentence, the source and the target language. $p_M(\mu)$ returns the probability of μ to be a problem in the real world. With p_T, we denote the distribution that is used in a benchmark. In the translation case, this is a selection of corpora that approximates p_M, not necessarily well. Then, instances can be described by some 'features', such as length, vocabulary, syntax, etc. The columns 'production' and 'verification' represent the way the examples in p_T are produced and verified. When humans appear in either or both of these columns, we have, as previously discussed, problems of extrapolation. The column 'proposed dimensions' suggests possible difficulty metrics. The column 'MT' indicates if AI is super/par/subhuman at the moment for p_M and p_T. Finally, the last column gives a reasonable guess of the extrapolation category for the domain, also as for today.

For the production phase, some problems require humans but not human cognition, such as diagnosis, driving, object recognition, etc. For verification, more and more problems also depend on humans, such as text, image, audio or video generation or transformation (e.g., translation), an area that has had significant progress recently. All these problems seem difficult to extrapolate beyond human level. Who is going to produce and verify problems that are harder than those that humans can conceive and/or solve?

There are already efforts for automated instance production and verification. For instance, Fig. 3 shows four synthetic images and the use of Fréchet Inception Distance (Heusel et al. 2017), to *verify* their quality, not without criticism (Kynkäänniemi et al. 2019; Barratt and Sharma 2018).

Bubble (FID = 63.5) | Baseball player (FID = 49.2) | Trumpet (FID = 100.4) | Park bench (FID = 80.3)

Figure 3: Some generated objects with poor (high) FID score. From (Kynkäänniemi et al. 2019).

The interesting thing of these proxies is that they rely on the *cognitive effort* of DL algorithms, such as Inception; another adversarial model is needed in order to verify these models automatically. Nonetheless, image generation quality for some applications depends on how well the image fools a human, so human verification is needed in the end. However, not all generative models must be like this. For instance, AI can generate a bridge or a molecule, with a non-anthropocentric quality metric. Overall, we must be very clear about what level and kind of cognitive effort (from a human or an AI system) is needed to produce and verify an example.

Multidimensional Space and Generality

The column 'instance features' in Table 1 represents the traits of the problem, but not the dimensions that should characterise the 'Universal' category in Fig. 2. A higher score in a dimension should represent a higher capability. As a result, the location of an agent in such a task space would be given by its *cognitive profile*, as a vector of capabilities. Inspired by similar representations from factor analysis and psychometric techniques, a spatial view has also been sug-

gested several times in AI (Bhatnagar and others 2017). One such a recent attempt (Osband and others 2019) is illustrated in Fig. 4 (left). In this example, however, some dimensions group a kind of problem (e.g., instances requiring memory), but others represent variations (such as Gaussian Noise). Are these dimensions actual 'abilities'? The answer dates back to Thurstone (1937) and other psychometricians: an agent is more able as it solves more difficult problems.

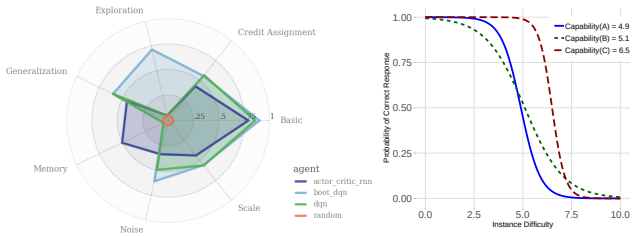

Figure 4: Left: four RL agents in a multidimensional space (Osband and others 2019). Right: Three *agent characteristic curves*, $\psi_{\pi,p_M}(h)$, and their areas (capabilities).

A multidimensional task space allows for different pathways of progress and different system profiles. The correspondence between each dimension and a well-defined difficulty function becomes crucial to fully understand what we are measuring. This is especially the case when we move from task-oriented evaluation to more ability-oriented evaluation (Hernández-Orallo 2017b) and the dimensions of more general competence in domains such as natural language (Yogatama and others 2019).

The identification of several dimensions for each domain suggests that profiles may be quite varied, as we have seen in some examples above. For instance, two systems with equal accuracy for an image classification benchmark can differ when images are rotated or blurred. The question is whether we can consider one system more general than the other. There are two ways to analyse this[2], intra-dimensionally and inter-dimensionally:

- Intra-dimensional generality refers to a system displaying consistent success for elements of low difficult in a particular dimension. For instance, curves in blue and red in Figure 4 (right) are steeper and hence ensure a more consistent (saturated) start of the curve, over the green curve, which is flatter. This notion of generality, related to the shape of an agent chracteristic curve, has been explored in (Martinez-Plumed and Hernandez-Orallo 2018).
- Inter-dimensional generality refers to a *balanced* result for all dimensions. For instance, if a machine vision system A is robust for medium levels of rotation and blur, while system B is very good at all levels of rotation but brittle for low levels of blur, then we consider A more general than B. The critical question to properly assess this balance for all dimensions depends on the magnitude of each dimension, and making them commensurable.

[2]Other approaches to generality take a more information-oriented perspective (Hernández-Orallo 2017a; Chollet 2019).

Identifying the dimensions of a domain is a very important decision that determines the domain space, and both the intra- and the inter-dimensional generality analysis. In psychological measurement, the dimensions are identified by factor analysis and other psychometric techniques. While this is not recommended in general, as the population of AI agents is very arbitrary (Hernández-Orallo 2017a), it can still be used for the analysis of data in competitions, especially when the distribution of participants is meaningful, or we want to establish *relative* measurement. When experiments are not used for deriving the dimensions, one has to think about natural skills or perturbations that could compose a meaningful set of categories for a task. For instance, in (Beyret et al. 2019) dimensions are identified as categories that cluster kinds of problems in animal cognition, even if this does not have a precise formal consideration, but based on accumulated knowledge in comparative cognition.

Conclusions

There are some recent initiatives, surveys and books covering the problem of AI evaluation, and giving comprehensive or partial views of the use of measurement theory in AI, such as (Hernández-Orallo 2017b; 2017a; Hernández-Orallo et al. 2017; Welty, Paritosh, and Aroyo 2019; Flach 2019b; 2019a; Avrim 2019). In this paper we have analysed the specific phenomenon of many benchmarks being discontinued as soon as AI systems reach a particular (human) level. This is an unusual situation in other disciplines, even if one understands that measurement instruments usually evolve at the same time progress takes place in the discipline. We did not throw away our speedometers the first time a train or a car ran faster than a human, or a cheetah.

We have argued for an extended view of AI evaluation, where humans should not be taken as reference but seen in a wider intelligence landscape (Bhatnagar and others 2017). Domains should be accompanied by dimensions that could be altered by non-cognitive distortions and cognitive modifications to make more difficult problems, in such a way that the space of evaluation stretches longer and wider than the trajectory that is defined by humans. We have seen some examples in Table 1.

The application of ideas from psychological measurement and psychometrics to AI, such as psychophysics (e.g., (Leibo and others 2018)) and item response theory (e.g., (Martínez-Plumed et al. 2019)), present many possibilities to curate data in benchmarks, determine difficulty and discrimation for their instances, and derive new instances with higher difficulties and discriminating power. The distinction between task-oriented evaluation and ability-oriented evaluation (Hernández-Orallo 2017b) should shed more light in organising the dimensions, e.g., in a hierarchical way.

Notwithstanding, the elephant in the room of AI evaluation is difficulty. Without a proper identification of the difficulty of each dimension, and its connection with resources, it will be hard –if not impossible– to derive meaningful scales for each dimension and, ultimately, units for each of them. I have claimed on many occasions that difficulty should be be derived from first principles, either computationally (Hernández-Orallo 2017a) or using the essence of

a task domain, as done here. A proper unit leading to a ratio measurement scale should be linked to difficulty: double difficulty should require double effort. Capability would have the same units too, being defined as the level of difficulty that can be achieved by a system (see Figure 4, right). The parallel between physical power and mental power, and their units in physics, is inspiring (Hernández-Orallo 2019), but the challenges for a theory of difficulty of AI are, at the moment, immeasurable.

Acknowledgements: This work was funded by the Future of Life Institute, FLI, under grant RFP2-152, and also supported by the EU (FEDER) and Spanish MINECO under RTI2018-094403-B-C32, and Generalitat Valenciana under PROMETEO/2019/098.

References

Amodei, D., and Hernandez, D. 2018. AI and compute. *https://blog. openai. com/aiand-compute*.

Avrim, G. 2019. Evaluation of artificial intelligence systems. Laboratoire Nationale d la Metrologie: https://www.lne.fr/en/testing/evaluation-artificial-intelligence-systems.

Barratt, S., and Sharma, R. 2018. A note on the inception score. *arXiv preprint arXiv:1801.01973*.

Beyret, B.; Hernández-Orallo, J.; Cheke, L.; Halina, M.; Shanahan, M.; and Crosby, M. 2019. The animal-ai environment: Training and testing animal-like artificial cognition. *arXiv preprint arXiv:1909.07483*.

Bhatnagar, S., et al. 2017. Mapping intelligence: requirements and possibilities. In *PTAI*, 117–135. Springer.

Carter, S., and Nielsen, M. 2017. Using artificial intelligence to augment human intelligence. *Distill*. https://distill.pub/2017/aia.

Cheng, J.; Huang, W.; Cao, S.; Yang, R.; Yang, W.; Yun, Z.; Wang, Z.; and Feng, Q. 2015. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PloS one* 10(10):e0140381.

Chollet, F. 2019. The measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Dodge, S., and Karam, L. 2017. A study and comparison of human and deep learning recognition performance under visual distortions. In *ICCCN*, 1–7. IEEE.

Flach, P. 2019a. Measurement theory for data science and AI: Modelling the skills of learning machines and developing standardised benchmark tests. Turing Institute, https://www.turing.ac.uk/research/research-projects/measurement-theory-data-science-and-ai.

Flach, P. 2019b. Performance evaluation in machine learning: The good, the bad, the ugly and the way forward. In *AAAI*.

George, D.; Lehrach, W.; Kansky, K.; Lázaro-Gredilla, M.; Laan, C.; Marthi, B.; Lou, X.; Meng, Z.; Liu, Y.; Wang, H.; et al. 2017. A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science* 358(6368):eaag2612.

Han, L. 2016. Machine translation evaluation resources and methods: A survey. *arXiv preprint arXiv:1605.04515*.

Hernández-Orallo, J.; Baroni, M.; Bieger, J.; Chmait, N.; Dowe, D. L.; Hofmann, K.; Martínez-Plumed, F.; Strannegård, C.; and Thórisson, K. R. 2017. A new AI evaluation cosmos: Ready to play the game? *AI Magazine* 38(3).

Hernández-Orallo, J. 2000. Beyond the Turing Test. *J. Logic, Language & Information* 9(4):447–466.

Hernández-Orallo, J. 2017a. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press.

Hernández-Orallo, J. 2017b. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review* 48(3):397–447.

Hernández-Orallo, J. 2019. Unbridled mental power. *Nature Physics* 15(1):106.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 6626–6637.

Joerger, M.; Jones, C.; Shuman, V.; and . 2019. Testing connected and automated vehicles (cavs). In *Road Vehicle Automation 5*. 197–206.

Kynkäänniemi, T.; Karras, T.; Laine, S.; Lehtinen, J.; and Aila, T. 2019. Improved precision and recall metric for assessing generative models. *arXiv preprint arXiv:1904.06991*.

Leibo, J. Z., et al. 2018. Psychlab: a psychology laboratory for deep reinforcement learning agents. *arXiv preprint arXiv:1801.08116*.

Martinez-Plumed, F., and Hernandez-Orallo, J. 2018. Dual indicators to analyse ai benchmarks: Difficulty, discrimination, ability and generality. *IEEE Transactions on Games*.

Martínez-Plumed, F.; Avin, S.; Brundage, M.; Dafoe, A.; hÉigeartaigh, S. Ó.; and Hernández-Orallo, J. 2018a. Accounting for the neglected dimensions of ai progress. *arXiv preprint arXiv:1806.00610*.

Martínez-Plumed, F.; Loe, B. S.; Flach, P.; O hEigeartaigh, S.; Vold, K.; and Hernández-Orallo, J. 2018b. The facets of artificial intelligence: A framework to track the evolution of AI. *IJCAI*.

Martínez-Plumed, F.; Prudêncio, R. B.; Martínez-Usó, A.; and Hernández-Orallo, J. 2019. Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial Intelligence* 271:18–42.

Masum, H.; Christensen, S.; and Oppacher, F. 2002. The Turing ratio: Metrics for open-ended tasks. In *Conf. on Genetic and Evolutionary Computation*, 973–980. Morgan Kaufmann.

McCorduck, P. 2004. *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. AK Peters Natick, MA.

McDermott, D. 2007. Level-headed. *Artificial Intelligence* 171(18):1183–1186.

Merler, M.; Ratha, N.; Feris, R. S.; and Smith, J. R. 2019. Diversity in faces. *arXiv preprint arXiv:1901.10436*.

Mueller, E. T. 1987. Daydreaming and computation. Technical report, TR, CSD-870017, Ph.D. dissertation, U. California.

Neyshabur, B.; Bhojanapalli, S.; McAllester, D.; and Srebro, N. 2017. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, 5947–5956.

Osband, I., et al. 2019. Behaviour suite for reinforcement learning. *arXiv preprint arXiv:1908.03568*.

Perez-Liebana, D.; Samothrakis, S.; Togelius, J.; Schaul, T.; and Lucas, S. M. 2016. General video game AI: Competition, challenges and opportunities. In *AAAI*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1(8).

Rajalingham, R.; Issa, E. B.; Bashivan, P.; Kar, K.; Schmidt, K.; and DiCarlo, J. J. 2018. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience* 38(33):7255–7269.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3):211–252.

Schlangen, D. 2019. Language tasks and language games: On methodology in current natural language processing research. *arXiv preprint arXiv:1908.10747*.

Shanahan, M. 2015. *The Technological Singularity*. MIT Press.

Shoham, Y.; Perrault, R.; Brynjolfsson, E.; Clark, J.; Manyika, J.; Niebles, J. C.; Lyons, T.; Etchemendy, J.; Grosz, B.; and Bauer, Z. 2018. The ai index 2018 annual report. *AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA*.

Shoham, Y. 2017. Towards the AI index. *AI Magazine* 38(4):71–77.

Silver, D., et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484.

Stern, R.; Sturtevant, N.; Felner, A.; Koenig, S.; et al. 2019. Multi-agent pathfinding: Definitions, variants, and benchmarks. *arXiv preprint arXiv:1906.08291*.

Sutton, R. S. 1996. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances in neural information processing systems*, 1038–1044.

Thurstone, L. 1937. Ability, motivation, and speed. *Psychometrika* 2(4):249–254.

Tucker, A. D.; Anderljung, M.; and Dafoe, A. 2020. Social and governance implications of improved data efficiency. *arXiv preprint arXiv:2001.05068*.

Van Seijen, H.; Fatemi, M.; Romoff, J.; Laroche, R.; Barnes, T.; and Tsang, J. 2017. Hybrid reward architecture for reinforcement learning. In *NIPS*, 5392–5402.

Vardi, M. Y. 2015. Human or machine? Response. *Communications of the ACM* 58(4):8–8.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Welty, C.; Paritosh, P.; and Aroyo, L. 2019. Metrology for AI: From benchmarks to instruments. *arXiv preprint arXiv:1911.01875*.

Yogatama, D., et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.