On Broken Yardsticks and Measurement Scales

José Hernández-Orallo (jorallo@upv.es)
vrAIn, Universitat Politècnica de València, Valencia (www.upv.es)
Also associate at Leverhulme Centre for the Future of Intelligence, Cambridge (lcfi.ac.uk)

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

Meta-Eval@AAAI2020, New York, Feb 8, 2020

SUPERHUMAN: BREAKING THE YARDSTICK!

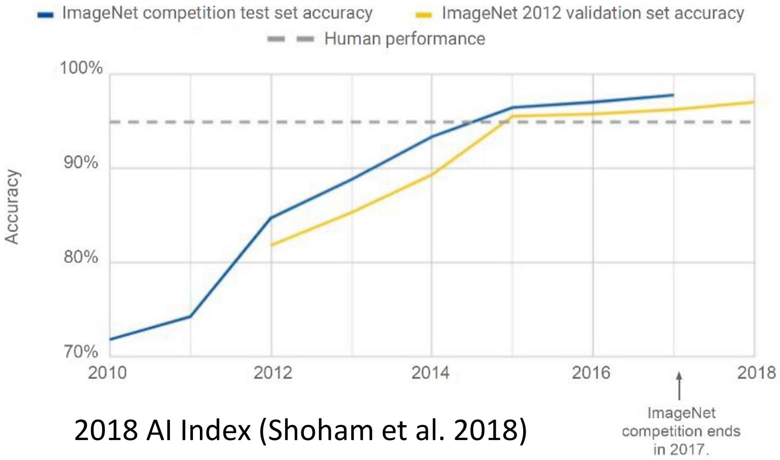
- Superhuman level is now reached for many tasks:
  - Can we meaningfully extrapolate beyond that level?
    - What is 999,990 points (HRA) in Pac-Man?
      - (Average human: 15,693, best human: 266,330)
    - Meaningless!
      - No worries, we build another benchmark
- In other tasks what does superhuman mean?
  - Superhuman translation?
    - Shouldn't we need humans to determine this?

AI evaluation suffers a moving target phenomenon:
tasks are replaced, more human effort needed

To Boldly Go To Humanity and Beyond!

- Beyond human performance
  - A 'challenge-solve-and-replace' evaluation dynamics (Schlangen 2019),
  - A 'dataset-solve-and-patch' adversarial benchmark coevolution (Zellers et al. 2019)
- Can we keep the benchmarks?
  - What's better-than-human Imagenet performance?
    - Is 97% improvement over 95% as relevant as 95% over 93%?
      - Is the magnitude meaningful?
      - Is extrapolation possible?

CIFAR10 → CIFAR100,
SQuAD1.1 → SQuAD2.0,
GLUE → SUPERGLUE,
Starcraft → Starcraft II



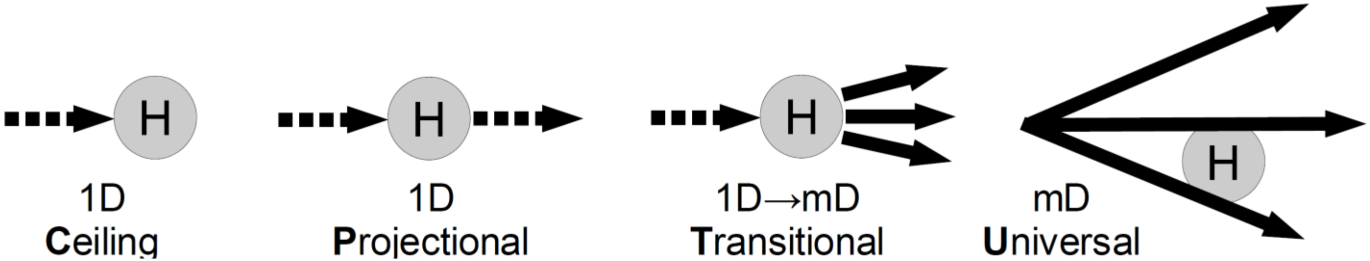
2018 AI Index (Shoham et al. 2018)

The Moving Target: Five Possible Causes

- Causes of this ‘challenge-solve-and-replace’ phenomenon
  - “AI effect” (McCorduck 2004): whenever something is automated, it’s **not intelligence** any more!
  - “Superhuman abyss”: once AI reaches superhuman level for a given task, there are many **arbitrary and unjustified extensions.**
  - “Resource neglect”: breakthroughs are obtained with huge **resources** in terms of data, compute, supervision and other internalities/externalities.
  - “Specialisation drift”: tendency of AI researchers to **specialise** to a particular task, or to overfit to a benchmark (Goodhart’s law, reproducibility).
  - “Cognitive-judge problem”: manual or automatic cognitive effort is needed to **produce and verify** instances (change distribution rather than make it harder).

EXTENSIBLE YARDSTICKS: EXTRAPOLATION POSSIBILITIES

- The 'Ceiling' (C) category sets humans 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., adding Gaussian noise to ImageNet, Dodge and Karam 2017).
- The 'Universal' (U) defines a (multidimensional) space from the very conception of the task (e.g., brain cancer diagnosis).



Domain Diversity: Unified Analysis

- Characterising all benchmarks:
  - Mother (problem) distribution p_M vs test (benchmark) distribution p_T
    - Naïve to assume they are equal
  - Instance (Meta-)Features
    - High-level features: type of objects in an image, text language, etc.
  - Dimensions (selection or combination of meta-features)
    - Mapped to difficulty metrics: contrast, no. objects or words, etc.
  - Production of instances
    - Collecting form the physical world or from human effort?
  - Verification of instances
    - Automatically, human judges, adversarially?

UNIFIED ANALYSIS: COMPARING DOMAINS

- Examples:

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

- Preliminary and non-exhaustive table.

Producing and Verifying Instances: Cognitive Effort

- Producing more difficulty instances. Types of distortions:
  - Psychophysics or simple distortions (e.g. noise)
  - Cognitive distortions:
    - Humans introducing distractors in a text
    - A generator creating modifications of existing instance: e.g., variations of a sentence
    - A generator creates completely new synthetic images:

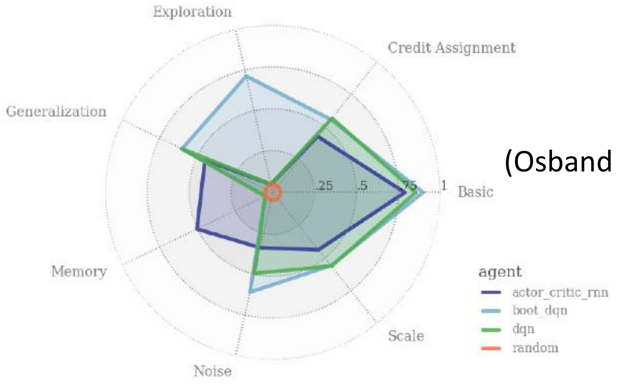
Bubble (FID = 63.5)	Baseball player (FID = 49.2)	Trumpet (FID = 100.4)	Park bench (FID = 80.3)

high FID is worse (Kynkäänniemi et al. 2019)

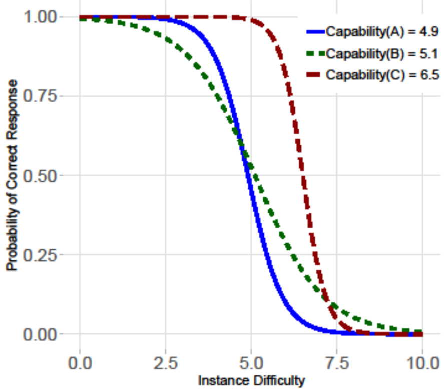
- Verifying them:
  - Fréchet Inception Distance not always accurate.
  - Relying of humans to check them (crowdsourcing)

MULTIDIMENSIONAL SPACES: INTER/INTRA-DIMENSIONAL GENERALITY

- The dimensions of difficulty make up a space:



(Osband et al. 2019)

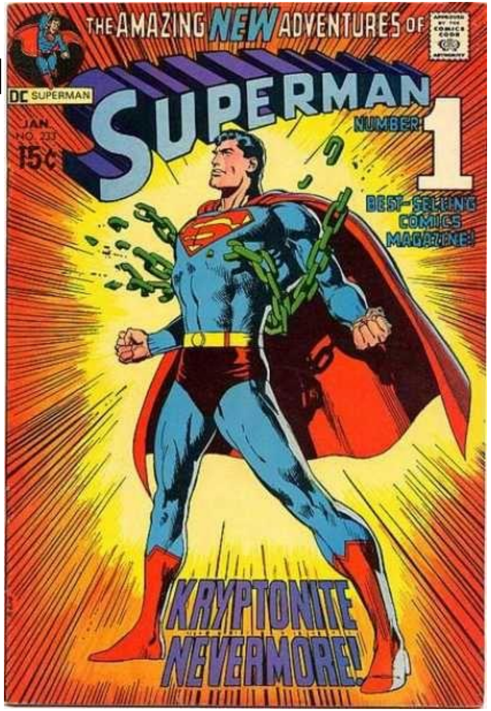


- Types of generality:
  - Inter-dimensional generality: balanced result for all dimensions: similar levels of rotation and blur.
  - Intra-dimensional generality: blue and red are steeper and hence ensure a more consistent (saturated) start of the curve, over the green curve.

Conclusions and Open Questions

- Superhuman performance breaks yardsticks that took humans as a ceiling or with instances produced and verified by humans.
  - Moving target issues, extrapolation issues, magnitudes, etc.
- The dimensions of difficulty allow for extrapolations, where humans are points in this space.
  - Commensurability issues
    - How do we choose the difficulty metrics?

We need a difficulty theory for AI

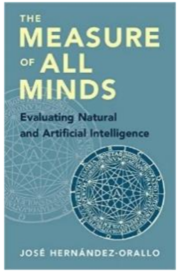


Ongoing Debates and Initiatives: Let's Work Together!

- It's getting momentum!
  - Moving from task-oriented to ability-oriented measurement (Hernández-Orallo 2017a, Cambridge University Press, 2017b, AIReviews)
  - Mapping the whole landscape of intelligence (Bhatnagar et al. 2017, PTAI)
  - Psychophysics in DRL benchmarks (Leibo 2018, arxiv)
  - Item Response Theory for ML/AI evaluation (Martínez-Plumed et al. 2019, AIJ)
  - Challenge-solve-and-replace evaluation dynamics (Schlangen 2019, arxiv)
  - Measurement theory for data science and AI at the Turing (Flach 2019, AAAI, https://www.turing.ac.uk/research/research-projects/measurement-theory-data-science-and-ai)
  - Multidimensional approach (Osband et al. 2019, arxiv).
  - Metrology for AI (Welty et al. 2019, arxiv).
  - Units of measurement (Hernández-Orallo 2019, Nature Physics)
  - EC's AI Collaboratory (Martínez-Plumed et al. 2020, ECAI): aicollaboratory.org

THANK YOU!

- Other Talks (http://josephorallo.webs.upv.es/)
  - The What and How of AI Evaluation
  - Diversity Unites Intelligence: Measuring Generality
  - Measuring A(G)I Right: Some Theoretical and Practical Considerations
  - Natural and Artificial Intelligence: Measures, Maps and Taxonomies
  - The Mythical Human-Level Machine Intelligence
- Book (http://allminds.org):
  - The Measure of All Minds: Evaluating Natural and Artificial Intelligence, Cambridge 2017



- Other Events:
  - epAI (Evaluating progress in AI, at ECAI, June 2020)
    - http://dmip.webs.upv.es/EPAI2020/