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)

Meta-Eval@AAAI2020, New York, Feb 8, 2020
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*
Beyond human performance
- A ‘challenge-solve-and-replace’ evaluation dynamics (Schlangen 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
  - “Al 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).
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:

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

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:

- 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:

- 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.

(Osband et al. 2019)
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)
  - 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/