MEASURING A(G)I RIGHT: SOME THEORETICAL AND PRACTICAL CONSIDERATIONS

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

DeepMind, London, UK, 29 June 2018
Are we measuring the right things in AI?
MEASURING AI SUCCESS TASK BY TASK: WE ARE PROGRESSING!

- **AI INDEX**

  - Object Detection, LSVRC Competition
  - Question Answering, SQuAD v1.1
  - Speech Recognition, Switchboard HUB5’00
  - Visual Question Answering, VQA 1.0

  ![Images of AI progress charts](https://www.eff.org/ai/metrics)

- **Tegmark’s “Life 3.0”**

  ![Tegmark's Life 3.0 diagram](https://www.eff.org/ai/metrics)
Measuring AI Success Task by Task: In Many Areas!

- **Specific (task-oriented) AI systems**
  - Prediction and estimation
  - Machine translation, information retrieval, summarisation
  - Robotic navigation
  - Knowledge-based assistants
  - Driverless vehicles
  - Planning and scheduling
  - Automated deduction
  - Game playing

PR: computer vision, speech recognition, etc.
MEASURING AI SUCCESS TASK BY TASK: COMPETITIONS FLOURISH!

Specific domain evaluation settings:

- CADE ATP System Competition → PROBLEM BENCHMARKS
- Termination Competition → PROBLEM BENCHMARKS
- The reinforcement learning competition → PROBLEM BENCHMARKS
- Program synthesis (Syntax-guided synthesis) → PROBLEM BENCHMARKS
- Loebner Prize → HUMAN DISCRIMINATION
- Robocup and FIRA (robot football/soccer) → PEER CONFRONTATION
- International Aerial Robotics Competition (pilotless aircraft) → PROBLEM BENCHMARKS
- DARPA driverless cars, Cyber Grand Challenge, Rescue Robotics → PROBLEM BENCHMARKS
- The planning competition → PROBLEM BENCHMARKS
- General game playing AAAI competition → PEER CONFRONTATION
- BotPrize (videogame player) contest → HUMAN DISCRIMINATION
- World Computer Chess Championship → PEER CONFRONTATION
- Computer Olympiad → PEER CONFRONTATION
- Annual Computer Poker Competition → PEER CONFRONTATION
- Trading agent competition → PEER CONFRONTATION
- Robo Chat Challenge → HUMAN DISCRIMINATION
- UCI repository, PRTools, or KEEL dataset repository → PROBLEM BENCHMARKS
- KDD-cup challenges and ML kaggle competitions → PROBLEM BENCHMARKS
- Machine translation corpora: Europarl, SE times corpus, the euromatrix, Tenjinno competitions... → PROBLEM BENCHMARKS
- NLP corpora: linguistic data consortium, ... → PROBLEM BENCHMARKS
- Warlight AI Challenge → PEER CONFRONTATION
- The Arcade Learning Environment → PROBLEM BENCHMARKS
- Pathfinding benchmarks (gridworld domains) → PROBLEM BENCHMARKS
- Genetic programming benchmarks → PROBLEM BENCHMARKS
- CAPTCHAs → HUMAN DISCRIMINATION
- Graphics Turing Test → HUMAN DISCRIMINATION
- FIRA HuroCup humanoid robot competitions → PROBLEM BENCHMARKS
- ...
This is still narrow:

- **Too much focus on given tasks**
  - Variations and artificial tasks are said not to be realistic or purposeful.

- **Too much focus on the final result**
  - Even transfer or curriculum learning look at the end of the development.

- **Too much focus on performance**
  - Teams aim for and papers designed to the test. At whatever cost!

- **Too much focus on specific tasks**
  - Divide-and-conquer AI philosophy. Two systems better than one?

- **Too much focus on humans**
  - As a reference or as an automation goal.
AI EVALUATION PLATFORMS: MORE FLEXIBLE

- These platforms make diverse task generation easier:
  - Facebook’s bAbi
  - Arcade Learning Env. (Atari)
  - Video Game Definition Language
  - OpenAI Gym
  - Microsoft’s Project Malmo
  - DeepMind Lab
  - DeepMind PsychLab
  - Mujoco
  - Facebook’s TorchCraft
  - Facebook’s CommAI

- But how is diversity and complexity created meaningfully?
  - Some tasks are created and we end up realising they are too easy or too hard for current AI. Moving targets?

Humans as a Reference: What Beyond?

- If a superhuman result is reached:
  - Was automation the only goal?
  - What’s the economy of getting better?
    - Can we quantify 1% better performance?
  - Is the task well-defined beyond humans?
    - Super-human perception?
    - Super-human translation?
- The task is replaced by a more difficult or challenging one:

https://www.eff.org/ai/metrics

The task is “solved”. So what are they doing?
### Neglected Dimensions! Focus on Resources

- **AI’s goal:** not really to automate tasks but to make them more efficient!
- **Many other resources (other than performance):**

<table>
<thead>
<tr>
<th>Resource</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>rd</td>
<td>Data</td>
<td>All kinds of data (unsupervised, supervised, queries, measurements). A self-driving car needs online traffic information.</td>
</tr>
<tr>
<td>rk</td>
<td>Knowledge</td>
<td>Rules, constraints, bias, utility functions, etc., that are required. A spam filter requires the cost matrix from the user.</td>
</tr>
<tr>
<td>rs</td>
<td>Software</td>
<td>Main algorithm, associated libraries, operating system, etc. A planner uses a SAT solver.</td>
</tr>
<tr>
<td>rh</td>
<td>Hardware</td>
<td>Computer hardware, sensors, actuators, motors, batteries, etc. A drone needs a 3D radar for operation.</td>
</tr>
<tr>
<td>rm</td>
<td>Manipulation</td>
<td>Manual (human-operated) intervention through assistance. A robot needs to be manually re-calibrated.</td>
</tr>
<tr>
<td>rc</td>
<td>Computation</td>
<td>Computational resources (CPU, GPU usage) of all the components. A nearest neighbor classifier computes all distances.</td>
</tr>
<tr>
<td>rm</td>
<td>Network</td>
<td>Communication resources (Internet, swarm synchronisation, distribution). An automated delivery system connects all drones.</td>
</tr>
<tr>
<td>rt</td>
<td>Time</td>
<td>Calendar (physical) time needed: waiting/night times, iteration cycles. A PA requires cyclical data (weeks) to find patterns.</td>
</tr>
</tbody>
</table>

---

*Resources that frequently appear (more or less explicitly) in AI systems*

Neglected Dimensions! Focus on Utility

- Reduce to a utility function (including performance).
  - That’s (partly) what makes a product innovative and successful!

![A schematic representation of different stages where resources might be required.](https://arxiv.org/abs/1806.00610)
The use of resources depends on many factors, but with all the dimensions we can see where the pareto-fronts are.

Go (left) and ALE (right). Research gradient evolution from 2013 to 2018 represented with a segmented grey arrow.

GENERAL-PURPOSE AI SYSTEMS: WHAT TO MEASURE HERE?

- How to evaluate general-purpose systems and cognitive components?

Cognitive robots

Pets, animats and other artificial companions

Agents, avatars, chatbots

Web-bots, Smartbots, Security bots...

Smart environments

Intelligent assistants
Evaluating General-Purpose AI: Is It Meaningful?

Intelligence is a subjective phenomenon. No-free-lunch theorems, multiple intelligences, narrow AI

Artificial systems: by conception, we can design a system to be good at A, C and I, and very bad at all the rest.

Non-human animals: environments, morphology, physiology and (co-)evolution creates some structure here.

Humans: strong correlation between cognitive tasks and abilities: general intelligence.

Intelligence is a convergent phenomenon. The positive manifold, g/G factors, Solomonoff prediction, AGI
EVALUATING GENERAL-PURPOSE AI: WHAT TESTS?

- The Turing Test?
  - and its myriad variants?
  - We moved “Beyond the Turing Test” two decades ago!
- It still has a strong influence on the narratives of AI evaluation and the future of AI:
  - “Mythical Turing Test” (Sloman, 2014):
    - Mythical human-level machine intelligence!

A red herring for general-purpose AI!

EVALUATING GENERAL-PURPOSE AI: WHAT TESTS?

- More comprehensive?
  - ARISTO (Allen Institute for AI) : College science exams
  - Winograd Schema Challenge : Questions targeting understanding.
  - Weston et al. “AI-Complete Question Answering” (bAbI)
  - CLEVR : Relations over visual objects

Now AI is superhuman on most of them!
(e.g., https://arxiv.org/pdf/1706.01427.pdf)

BEWARE: AI-Completeness claimed before Calculation, Chess, Go, Turing test, ...
Evaluting General-Purpose AI: What Tests?

- What about psychometric tests or animal tests in AI?
  - These tests are used for humans everywhere!
- In 2003, Sanghi & Dowe: simple program passed many IQ tests.
- This has not been a deterrent!
  - Psychometric AI (Bringsjord and Schimanski 2003):
    - An “agent is intelligent if and only if it excels at all established, validated tests of intelligence”.
  - Detterman, editor of the Intelligence Journal, posed “A challenge to Watson” (Detterman 2011)
    - 2nd level to “be truly intelligent”: tests not seen beforehand.
      - Response: “IQ tests are not for machines, yet” (Dowe & Hernandez-Orallo 2012)
EVALUATING GENERAL-PURPOSE AI: WHAT TESTS?

What about developmental tests (or tests for children)?

- Developmental robotics:
  - Battery of tests (Sinapov, Stoytchev, Schenk 2010-13)

- Cognitive architectures:
  - Newell “test” (Anderson and Lebiere 2003)
  - “Cognitive Decathlon” (Mueller 2007).


<table>
<thead>
<tr>
<th>Category</th>
<th>Level</th>
<th>PEBL</th>
<th>CHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Invariant Object Identification</td>
<td>Yes</td>
<td>Gv</td>
<td></td>
</tr>
<tr>
<td>Object ID: Size Discrimination</td>
<td>Yes</td>
<td>Gv</td>
<td></td>
</tr>
<tr>
<td>Vision</td>
<td>Object ID: With Rotation</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Vision</td>
<td>Object ID: Relations</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Vision</td>
<td>Visual Action/Event Recognition</td>
<td>No</td>
<td>Gv/Gl</td>
</tr>
<tr>
<td>Search</td>
<td>Simple Navigation</td>
<td>Yes</td>
<td>Gv</td>
</tr>
<tr>
<td>Search</td>
<td>Visual Search</td>
<td>Yes</td>
<td>Gv/Gs</td>
</tr>
<tr>
<td>Search</td>
<td>Travelling Salesman Problem</td>
<td>Yes</td>
<td>Gv/Gs/Gl</td>
</tr>
<tr>
<td>Search</td>
<td>Embodied Search</td>
<td>No</td>
<td>Gv/Gs/Gl</td>
</tr>
<tr>
<td>Search</td>
<td>Reinforcement Learning</td>
<td>Yes</td>
<td>Gv/Gs/Gl/Gf/Gm</td>
</tr>
<tr>
<td>Manual</td>
<td>Motor Mimicry</td>
<td>No</td>
<td>Gm/Gv</td>
</tr>
<tr>
<td>Control and Learning</td>
<td>Simple (One-Hand) Manipulation</td>
<td>Yes</td>
<td>Gm</td>
</tr>
<tr>
<td>Control and Learning</td>
<td>Two-Hand Manipulation</td>
<td>No</td>
<td>Gm/Gv</td>
</tr>
<tr>
<td>Control and Learning</td>
<td>Device Mimicry</td>
<td>Yes</td>
<td>Gm/Gv</td>
</tr>
<tr>
<td>Control and Learning</td>
<td>Intention Mimicry</td>
<td>No</td>
<td>Gm/Gv</td>
</tr>
<tr>
<td>Knowledge Learning</td>
<td>Episodic Recognition Memory</td>
<td>No</td>
<td>Gf/Gm?</td>
</tr>
<tr>
<td>Knowledge Learning</td>
<td>Semantic Memory/Categorization</td>
<td>No</td>
<td>Gf/Gm?</td>
</tr>
<tr>
<td>Language and Concept</td>
<td>Object-Noun Mapping</td>
<td>No</td>
<td>Gc/Gl</td>
</tr>
<tr>
<td>Language and Concept</td>
<td>Property-Adjective</td>
<td>No</td>
<td>Gc/Gl</td>
</tr>
<tr>
<td>Language and Concept</td>
<td>Relation-Preposition</td>
<td>No</td>
<td>Gc/Gl</td>
</tr>
<tr>
<td>Language and Concept</td>
<td>Action-Verb</td>
<td>No</td>
<td>Gc/Gl</td>
</tr>
<tr>
<td>Language and Concept</td>
<td>Relational Verb-Action</td>
<td>No</td>
<td>Gc/Gl</td>
</tr>
<tr>
<td>Simple Motor Control</td>
<td>Eye Movements</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>Simple Motor Control</td>
<td>Aimed Manual Movements</td>
<td>Yes</td>
<td>-</td>
</tr>
</tbody>
</table>
Evaluating General-Purpose AI: New Foundation

- Adapting tests between disciplines (AI, psychometrics, comparative psychology) is problematic:
  - Test from one group only valid and reliable for the original group.
    - No measurement invariance.
    - Not necessary and/or not sufficient for the ability.
  - Machines and hybrids represent a new population.
  - But machines and hybrids are also an opportunity to understand what cognitive tasks and cognitive abilities really are.

All cognitive tasks or environments $M$.

- $M$ only makes sense with a probability measure $p$ over all tasks $\mu \in M$.
- An agent $\pi$ is selected or designed for this $\langle M, p \rangle$.

$$\Psi(\pi, M, p) \triangleq \sum_{\mu \in M} p(\mu) \cdot R(\pi, \mu)$$

- If $M$ is infinite and diverse policies are acquired or learnt, not hardwired.
  - But who sets $\langle M, p \rangle$?
    - In biology, natural selection (physical world, co-evolution, social environments).
    - In AI, applications (narrow or more robust/adaptable to changes).
In a RL setting choosing a universal distribution \( p(\mu) = 2^{-K_U(\mu)} \) we get the so-called “Universal Intelligence” measure (Legg and Hutter 2007).

Proper formalisation of including all tasks, “generalising the C-test (Hernandez-Orallo 2000) from passive to active environments”.

Problems (pointed out by many: Hibbard 2009, Hernandez-Orallo & Dowe 2010):

- The probability distribution on \( M \) is not computable.
- Time/speed is not considered for the environment or agent.
- Most environments are not really discriminating (hells/heavens).
- The mass of the probability measure goes to just a few environments.

Legg and Hutter’s measure is “relative” (Leike & Hutter 2015), a schema for tasks, instantiated by a particular choice of the reference \( U \).
THE SPACE OF ALL POLICIES: AGENT CHARACTERISTIC CURVES

- Instead of the (Kolmogorov) complexity of the description of a task:
  - We look at the policy, the solution, and its complexity/resources.
  - The resources or computation it needs: this is the difficulty of the task.
  - For instance, \( K_{t_U}(x) \triangleq \min_p \frac{L_S(p)}{U(p) = x} \) with \( L_S(p) \triangleq L(p) + \log S(p) \)

- “Agent characteristic curves” (ACCs), expected response \( \Psi \) against difficulty:

- Agent resources can be used in cooperative/competitive scenarios (e.g., games)

---

Hernández-Orallo J. C-tests revisited: back and forth with complexity. AGI 2015
THE SPACE OF ALL POLICIES: AGGREGATION

- Alternative formulations:

Direct: Items derive from the representational distribution
\[ p_M(\mu) \]
[universal, e.g. Legg and Hutter]

Indirect: Items derive from difficulty.
\[ p_H(h) \]
[uniform]

Indirect through policies: Items derive from policies, policies from difficulty.
\[ p_H(h) \]
[uniform]
\[ p_H(\pi|h) \]
[uniform]
\[ p_M(\mu|\pi) \]
[Kt universal]

Less subjective.

Generalising the C-test right

Range of difficulties
Diversity of solutions: actual cognitive diversity

Less dependent on the representational mechanism for policies (invariance theorem).
Are we measuring AI in the right way?
HOW TO MEASURE: REPRESENTATIONAL MEASUREMENT

- If we know the set of tasks and their relevance/probability:

\[ \Psi(\pi, M, p) \triangleq \sum_{\mu \in M} p(\mu) \cdot R(\pi, \mu) \]

- Sampling \( M \) using \( p \) is not the most efficient way of estimating this reliably:
  - Some tasks do not discriminate or discriminate negatively
    - Some may be too easy or too difficult.
  - Redundant tasks do not provide information and agents can specialise for them.
  - The tasks with highest \( p \) will be common and agents will specialise for them.

We have to sample and then reconstruct \( \Psi \):
Redundant tasks must have their weight recovered for \( \Psi \)
HOW TO MEASURE: REPRESENTATIONAL MEASUREMENT

- Information-driven sampling.
  - Related to importance sampling and stratified sampling.
  - Diversity-driven sampling:
    - Given a similarity, e.g., derived from a set of features
    - We need to sample on $M$ such that:
      - the accumulated mass on $p$ is high.
      - diversity has to be maximised.
  - Difficulty-driven sampling.
    - The idea is to choose a range of difficulties with high weight.
    - Difficulty is defined as function $h: M \rightarrow \mathbb{R}$.
      - $h(\mu)$ must be monotonically decreasing on $E_\pi[\Psi(\pi, \mu, p)]$
      - More informative difficulties are covered.
  - Adaptive sampling
    - Reuses the results so far to find the most informative instance.

Covering $p$ without sampling very similar exercises repeatedly, and correcting the results accordingly (e.g., cluster sampling).

The results below $h=5$ and above $h=15$ can be assumed to be known, so effort is focussed on the relevant range.
**How to Measure: Operational Measurement**

- In operational (or pragmatic) measurement, there is no \( \Psi \).
  - We don’t have a definition of what we’re measuring.
  - Some tasks/tests, are useful as predictors or correlators of behaviour.
    - For instance, high IQ in humans is negatively correlated with religiosity.
  - Usually this predictability or correlations are \( \text{wrt. a population.} \)
  - We keep those tasks that show more variability for that population.
    - We end up dropping those that are too easy or too hard.
  - Measurement becomes *populational*: the measure of agent A not only depends on the choice of tasks but on the other agents in the population!

This is an iterative process, but sometimes this is criticised as a circular process.
HOW TO MEASURE: OPERATIONAL MEASUREMENT

- Do the tasks have the same magnitude or relevance (commensurability)?
  - For dichotomous tasks (correct or not), this is less critical than for quantitative tasks (e.g., scores).
  - Usual approaches in AI (especially ML):
    - Scaling (using the mean and the variance, or using quantiles).
    - Dichotomise by using a threshold (e.g., human performance).
    - Compare or average ranks (similar to scaling using quantiles).

All these solutions have advantages and disadvantages, but they always include an important bias in the measurement.
HOW TO MEASURE: SCALES AND UNITS

- Populational measurement is rarely conformant to ratio scales.
  - Ordinal scale: comparisons < and > are meaningful. No cycles!!!
  - Quantiles are used instead, e.g., IQ (100 mean, 15 sd),
    - We cannot compare the values additively (interval scale) or multiplicatively (ratio scale).
- Cannot compare values between two different populations
  - No common unit.
    - But possible with the policy-general approach (Hernandez-Orallo 2018)
  - Problems of measurement invariance.

Can we use this psychometric approach in AI?
Does a population of AI agents or techniques make sense?

Hernández-Orallo, J. “Intelligence without measure” Nature Physics, to appear
Psychometric Approach: Factor Analysis

- Behavioural latent features identified:
  - Personality traits: e.g., big five.
  - Cognitive abilities: primary abilities, g factor, hierarchical models.

- Tensions between one-factor (general intelligence) and “multiple intelligences”, sorted out by hierarchical models (and other SEM models)
**Psychometric Approach: Item Response Theory**

- How can we understand/improve items in a test?
  - Item Response Theory:
    - Logistic models: difficulty $b$, discrimination $a$ and guessing $c$.

![Diagram showing the 3PL-model](image)

$$p(\theta) = c + \frac{1-c}{1+e^{-a(\theta-b)}}$$

**Result matrix**

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Tests</th>
</tr>
</thead>
</table>

**Proficiency (ability) as achievable difficulty**
ITEM RESPONSE THEORY: APPLICATION TO ML/AI

- 49 Atari games (ALE) and 40 techniques.
- 2PL models: difficulty and discrimination vs ability

A Measure of Generality: Disentangling General Intelligence

A fundamental question for:

- Human intelligence: positive manifold, g factor. General intelligence?
- Non-human animal intelligence: g and G factors for many species. Convergence?
- Artificial intelligence: general-purpose AI or AGI. What does the G in AGI mean?

Usual interpretation:

General intelligence is usually associated with competence for a wide range of cognitive tasks

This is wrong! Any system with limited resources cannot show competence for a wide range of cognitive tasks, independently of their difficulty!

<table>
<thead>
<tr>
<th></th>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
<th>$\mu_3$</th>
<th>$\mu_4$</th>
<th>$\mu_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_a$</td>
<td>0.85</td>
<td>0.75</td>
<td>0.80</td>
<td>0.85</td>
<td>0.75</td>
</tr>
<tr>
<td>$\pi_b$</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
A Measure of Generality: It’s all about difficulty

General intelligence must be seen as competence for a wide range of cognitive tasks up to a certain level of difficulty.

- **Definition**
  - Capability (Ψ), the area under the ACC: $\psi_j = \int_0^\infty \psi_j^h dh$
  - Expected difficulty given success:
    $$H_j = \mathbb{E}_i[h|A_{i,j} = 1] = \frac{m_j}{\psi_j}$$
    $$m_j = \int_0^\infty h \cdot \psi_j^h dh$$
  - Spread:
    $$z_j = \sqrt{2(H_j - \psi_j) \cdot \psi_j} = \sqrt{2m_j - \psi_j^2}$$
  - Generality:
    $$\gamma_j = \frac{1}{z_j} = \frac{1}{\sqrt{2m_j - \psi_j^2}}$$

Non-populational!
A Measure of Generality: Some Agent Characteristic Curves
A Measure of Generality: Examples with Humans

- Capability and generality are observables, applied to individuals, no models.
- We don’t assume any grouping of items into tests with ranging difficulties.
- Applicable to individual agents and small sets of tasks/items.

Example (joint work with B.S. Loe, 2018):
- Elithorn’s Perceptual Mazes: 496 participants (Amazon Turk).
- Intrinsic difficulty estimators (Buckingham et al. 1963, Davies & Davies 1965).
- We calculate the generalities for the 496 humans.
  - Correlation between spread (1/gen) and capability is -0.53.
- See relation to latent main (general) factor:
  - All data: one-factor loading: 0.46, prop. of variance: 0.23.
  - 1stQ of generality: 1-f loading: 0.65, prop. of variance: 0.46.
**A Measure of Generality: A Definition of AGI?**

- How can the G in AGI be properly defined? No AI populations!
  - We want to calculate the generality of one AI system.
  - Using the new measure of generality:
    - We could have very general systems, with low capability.
    - They could be AGI but far from humans: baby AGI, limited AGI.
  - All other things equal, it makes more sense to cover easy tasks. first.
- Link to resources and compute.
  - Measuring capability and generality and their growth.
  - Look at superintelligence in this context.
- Generality leads to measurement transitivity:
  - Task transitivity: If A solves T1, and T2 is easier that T1 then A solves T2.
  - Agent transitivity: If A solves T, and B is more able than A, then B solves T.
A Measure of Generality: Example with AI

- Example (joint work with F. Martinez-Plumed 2018)
  - ALE (Atari games) and GVGAI (General Video Game AI) benchmarks.
  - Progress has been made, but what about generality? Are systems more general?
CONCLUSIONS

- Are we measuring the right things in AI?
  - Too focused on performance and specialised tasks.
  - Many dimensions (data, compute, human overseeing, etc.) are neglected.
  - Many new benchmarks (ALE, GVGAI, ...) are said to evaluate more general-purpose AI, but why is it so?
  - Theoretical approaches for general-purpose AI possible, based on difficulty.

- Are we measuring AI right?
  - When the measure is not representational, many things are biased (selection, scalings, etc.) or inconsistent (incommensurability, units, etc.)
  - We can take a populational approach (for competitions) or adversarial cases.
  - Non-populational approaches (e.g., generality) require difficulty/resources.
ONGOING INITIATIVES

- AEGAP at ICML/IJCAI this year: [http://cadia.ru.is/workshops/aegap2018/](http://cadia.ru.is/workshops/aegap2018/)
  - And other events about measurement in AI:
- Generality and AGI Risks:
  - How does generality affect AGI safety, together with capability and resources?
- The Atlas of Intelligence:
  - Collection of maps comparing humans, non-human animals and AI systems.

THANK YOU!