THE WHAT AND HOW OF AI EVALUATION

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

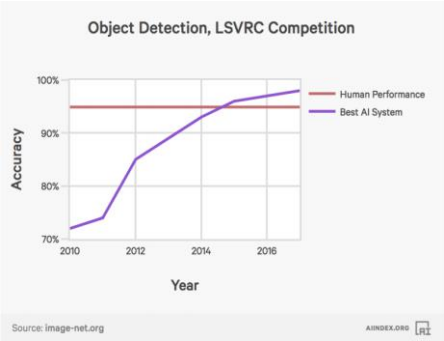
* With many thanks to Fernando Martínez-Plumed

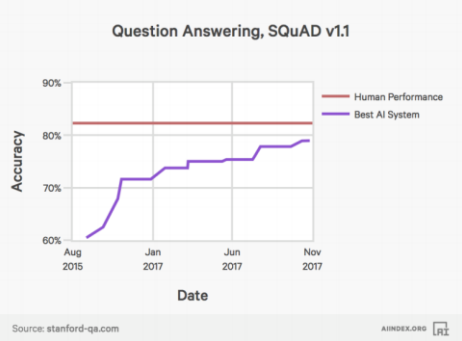
~~Winter~~ Early Spring
HUMAINT winter school on AI and its ethical, social, legal and economic impact
Joint Research Centre, European Commission
Seville, Spain, February 4-8th 2019

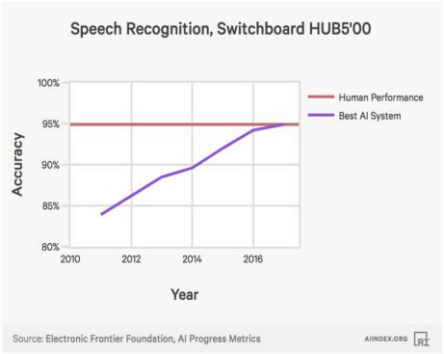
Are we measuring the right things in AI?

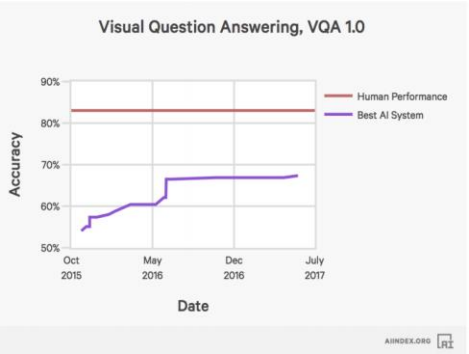
MEASURING AI SUCCESS TASK BY TASK: WE ARE PROGRESSING!

- AI INDEX









artificial intelligence index
2017 Annual Report

artificial intelligence index
2018 annual report

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

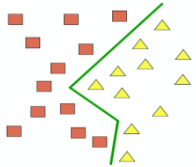
- Tegmark's "Life 3.0"

- Shoham at al. "AI Index Report" 2018 - http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf

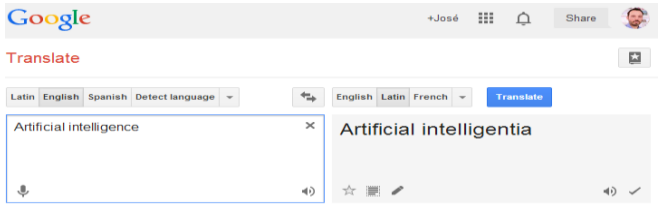
MEASURING AI SUCCESS TASK BY TASK: IN MANY AREAS!

- Specific (task-oriented) AI systems

PR: computer vision, speech recognition, etc.



Prediction and estimation

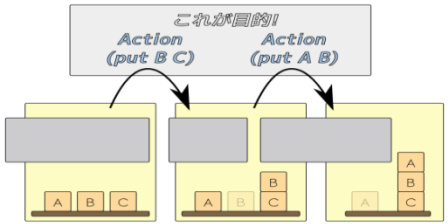


Machine translation, information retrieval, summarisation

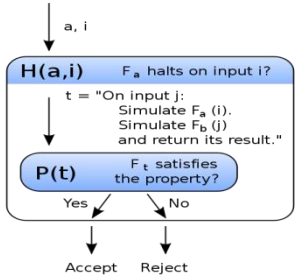
Robotic navigation

Knowledge-based assistants

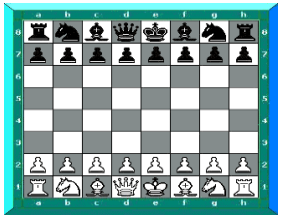
Driverless vehicles



Planning and scheduling



Automated deduction



Game playing

All images from wikicommons

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
  - …

Measuring AI Success Task by Task: Look under the carpet!

- This is still narrow:
  - Too much focus on fixed tasks (datasets or collections of tasks)
    - Generated variations are sometimes excluded (“they are not real” ☹).
  - Too much focus on specific tasks
    - Divide-and-conquer AI philosophy. Two systems better than one?
  - Too much focus on performance
    - Teams aim for and papers designed to the test. At whatever cost!
  - Too much focus on the final result
    - Even transfer or curriculum learning look at the *end* of the learning curves.
  - Too much focus on humans
    - As a reference or as an automation goal.

AI Evaluation Platforms: More flexible

- These platforms make it easier to create many tasks:
  - 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
  - Unity ML
- But how is diversity and complexity created meaningfully?
  - Too easy or too hard for current AI. Moving targets?

Malmo: "complexity gradient"

Bordes et al: "tasks of increasing difficulty"

Universe: "solve successively harder environments"

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue

Archive | Volume 540 | Issue 7633 | News | Article

NATURE | NEWS

عربي

Tech giants open virtual worlds to bevy of AI programs

Artificial-intelligence algorithms can learn a lot from playing immersive 3D video games.

Davide Castelvecchi

14 December 2016 | Corrected: 21 December 2016

PDF | Rights & Permissions

Matthew Tostevin/Reuters

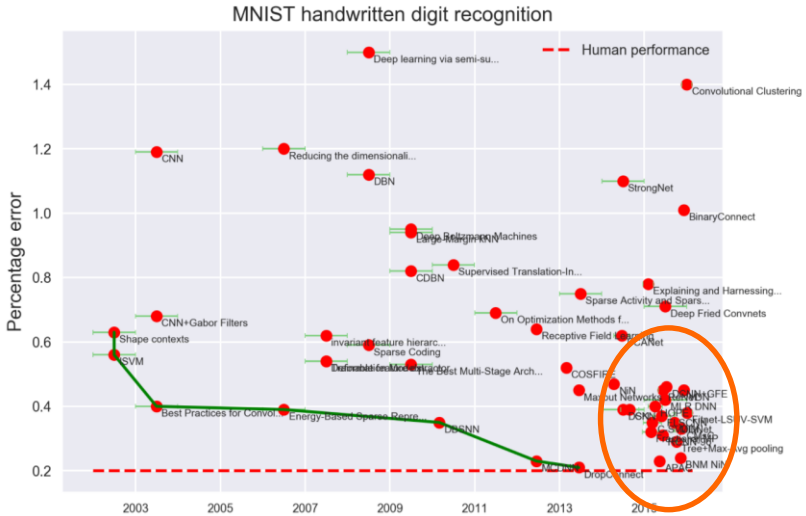
The *Minecraft* video game is popular with children; now a version is being used to test artificial-intelligence programs.

The *Minecraft* video game was familiar to José Hernández-Orallo long before he started using it for his own research. The computer

Hernández-Orallo, J., Baroni, M., Bieger, J., Chmait, N., Dowe, D.L., Hofmann, K., Martínez-Plumed, F., Strannegård, C. and Thórisson, K.R. "A New AI Evaluation Cosmos: Ready to Play the Game?" AI Magazine, (2017).

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

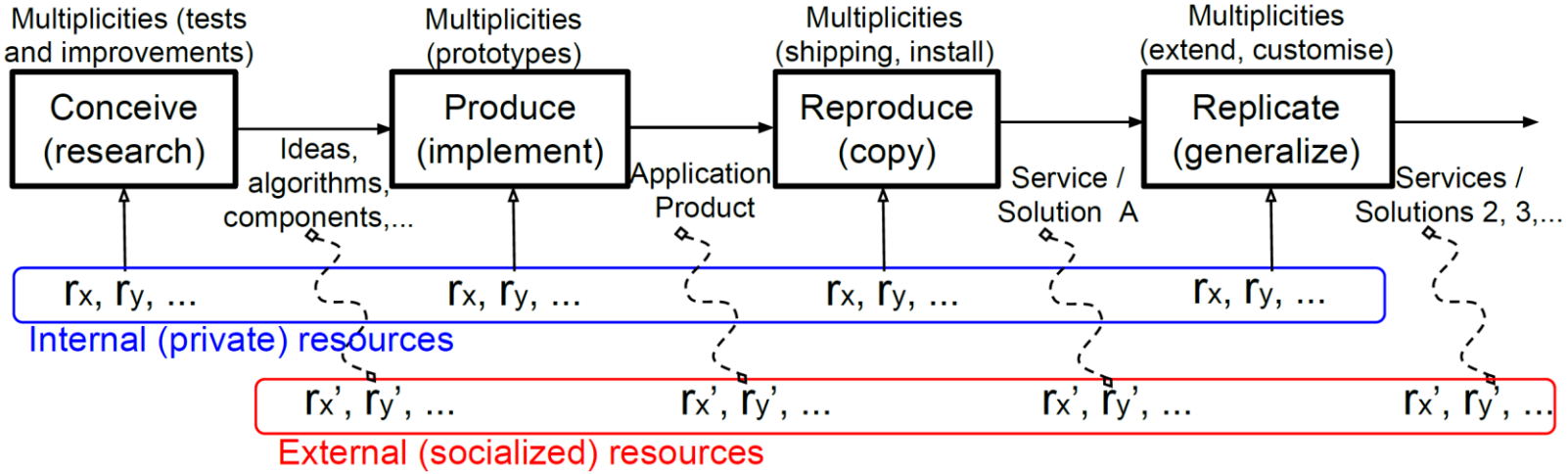
- AI's goal: not really to automate tasks but to make them more efficient!
  - Many other resources (other than performance):

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

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

Neglected Dimensions! Focus on Utility

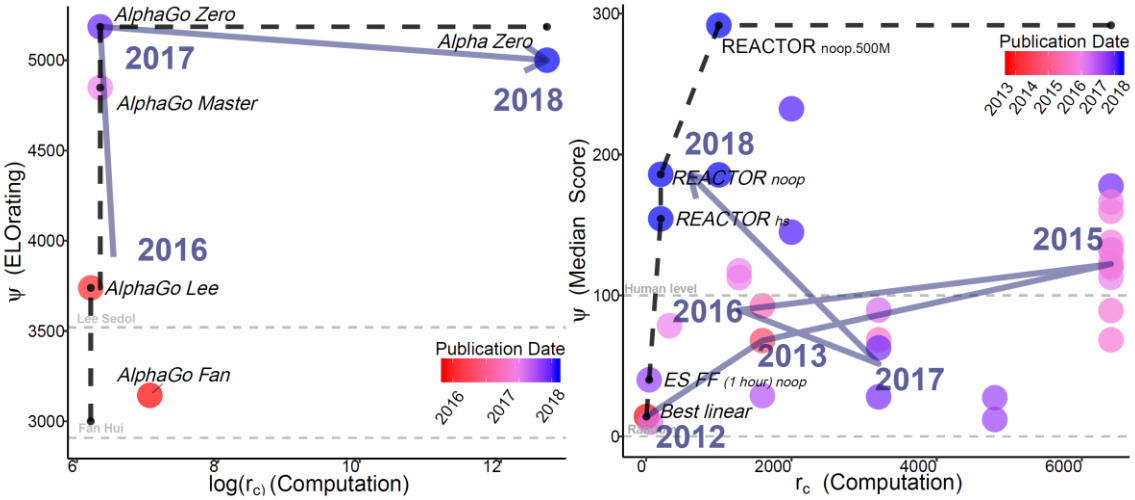
- Reduce all resources to a utility function (including performance).
  - That's (partly) what makes a product innovative and successful!



A schematic representation of different stages where resources might be required.

NEGLECTED DIMENSIONS! LET'S PLOT SOME OF THEM

- 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?

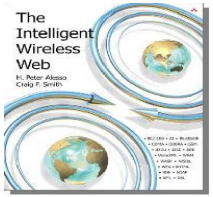
- How to evaluate general-purpose systems and cognitive components?

Cognitive robots

Pets, animats and other artificial companions

Smart environments

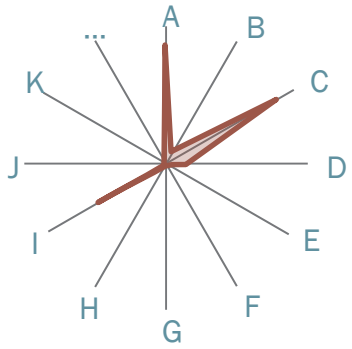
Agents, avatars, chatbots

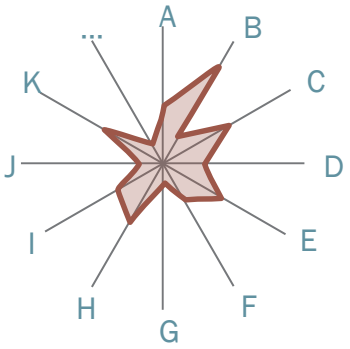


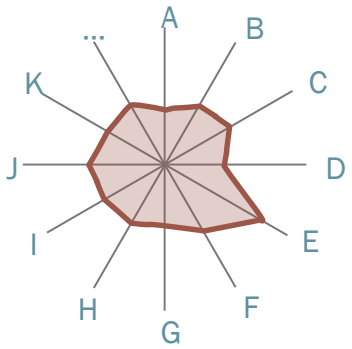
Web-bots, Smartbots, Security bots…

Intelligent assistants

EVALUATING GENERAL-PURPOSE AI: IS IT MEANINGFUL?







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

SPECIFIC

GENERAL

Intelligence is a convergent phenomenon. The positive manifold, g/G factors, Solomonoff prediction, AGI

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.

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!

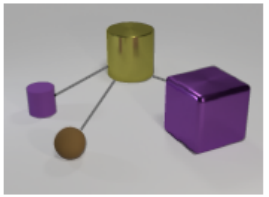
Hernández-Orallo, J. "Beyond the Turing Test" Journal of Logic, Language and Information, 2000.
Sloman, A. "Mythical Turing Test", http://www.cs.bham.ac.uk/research/projects/cogaff/misc/turing-test.html

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

Relational question:

Are there any rubber things that have the same size as the yellow metallic cylinder?

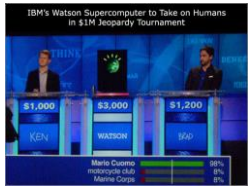


Now AI is superhuman on most of them!
(e.g., https://arxiv.org/pdf/1706.01427.pdf)

BEWARE: AI-Completeness claimed before
Calculation, Chess, Go, Turing test, …

Evaluating general-purpose AI: What Tests?

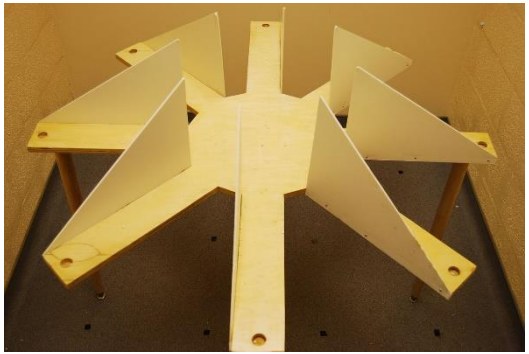
- What about psychometric tests or animal tests in AI?
  - These tests are used for humans everywhere!
- In 2003, Sanghi & Dowe: simple program passed many IQ tests.
  - This has not been a deterrent!
    - Psychometric AI (Bringsjord and Schmimanski 2003):
      - An "agent is intelligent if and only 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)



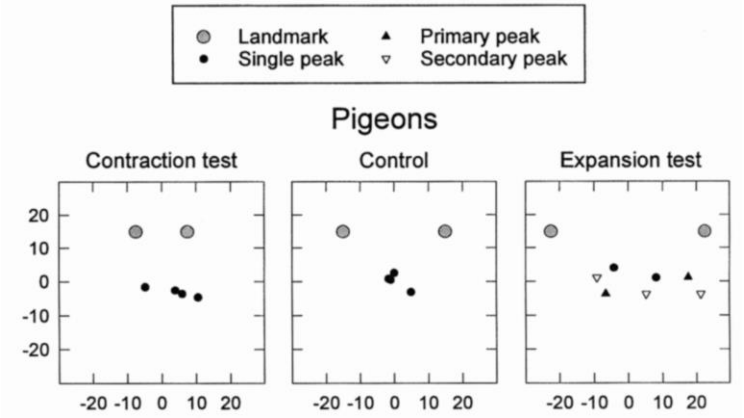
Sanghi, P., & Dowe, D. L. "A computer program capable of passing IQ tests", ICCS 2003
Bringsjord, S., & Schimanski, B. "What is artificial intelligence? Psychometric AI as an answer", IJCAI 2003.
Detterman "A challenge to Watson" *Intelligence Journal, 2011, and response* Dowe & Hernandez-Orallo "IQ tests are not for machines, yet"

Evaluating general-purpose AI: What Tests?

- What about tests from comparative cognition?
  - Mazes, A not B, Detour task, Landmark navigation, Middle cup, String pulling, Primate Cognition Test Battery, …



Radial arm maze



Spetch et al. 1997

Esther Hermann

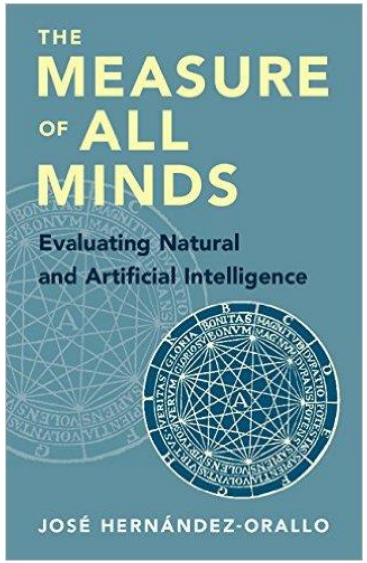
Evaluating general-purpose AI: What Tests?

- 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).
  - AGI: high-level competency areas (Adams et al. 2012), task breadth (Goertzel et al 2009, Rohrer 2010), robot preschool (Goertzel and Bugaj 2009).

		a taxonomy for cognitive architectures	a psychometric taxonomy (CHC)
Category	Level	PEBL	CHC
Vision	Invariant Object Identification	Yes	Gv
	Object ID: Size Discrimination	Yes	Gv
	Object ID: With Rotation	Yes	Gv
	Object ID: Relations	No	-
	Visual Action/Event Recognition	No	Gv|Gl
Search	Simple Navigation	Yes	Gv
	Visual Search	Yes	Gv|Gs
	Travelling Salesman Problem	Yes	Gv|Gs|Gl
	Embodied Search	No	Gv|Gs|Gl
	Reinforcement Learning	Yes	Gv|Gs|Gl|Gf|Gm
Manual Control and Learning	Motor Mimicry	No	Gm|Gv
	Simple (One-Hand) Manipulation	Yes	Gm|Gv
	Two-Hand Manipulation	No	Gm|Gv
	Device Mimicry	Yes	Gm|Gv
	Intention Mimicry	No	Gm|Gv
Knowledge Learning	Episodic Recognition Memory	No	Gl|Gm?
	Semantic Memory/Categorization	No	Gl|Gf|Gm?
Language and Concept Learning	Object-Noun Mapping	No	Gc|Gl
	Property-Adjective	No	Gc|Gl
	Relation-Preposition	No	Gc|Gl
	Action-Verb	No	Gc|Gl
	Relational Verb-Action	No	Gc|Gl
Simple Motor Control	Eye Movements	No	-
	Aimed Manual Movements	Yes	-

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.
    - An opportunity to understand what cognitive **tasks** and cognitive **abilities** really are.



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 Ψ:
Redundant tasks must have their weight recovered for Ψ

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
  - Quantitative tasks. E.g., can we add breakout scores with invaders 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.

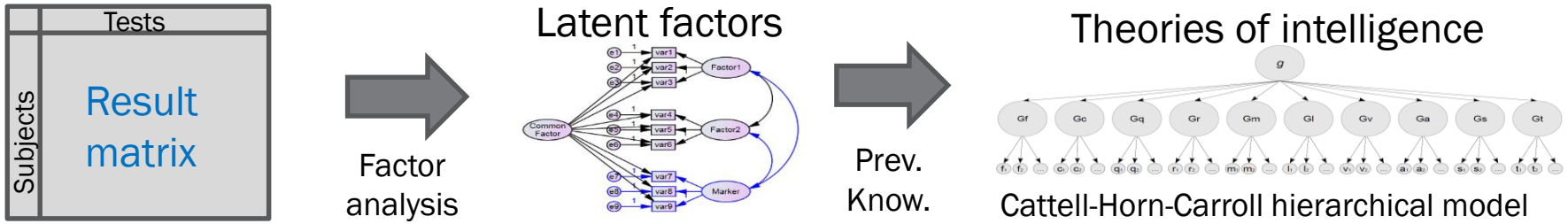
How to Measure: Scales and Units

- Populational measurement is rarely conformant to ratio scales.
  - Quantiles (ordinal scale) are used instead, e.g., IQ (100 mean, 15 sd),
    - We cannot compare 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 2019)
    - Problems of measurement invariance.

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

PSYCHOMETRIC APPROACH: FACTOR ANALYSIS

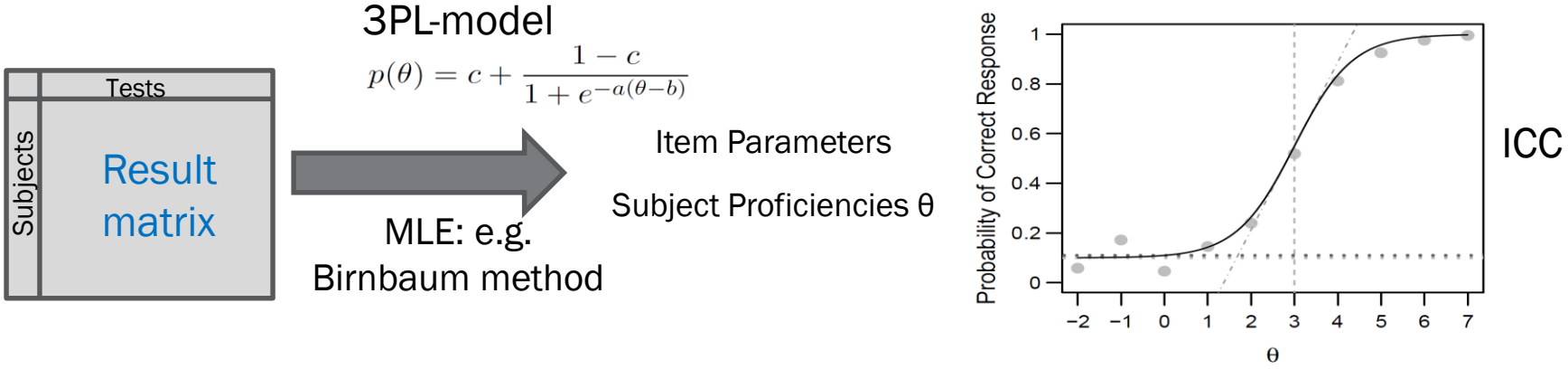
- Behavioural latent features identified:
  - Personality traits: e.g., big five.
  - Cognitive abilities: primary abilities, g factor, hierarchical models.



  - Tensions between one-factor (general intelligence) and “multiple intelligences”, sorted out by hierarchical models (and other SEM models)

Psychometric Approach: Item Response Theory

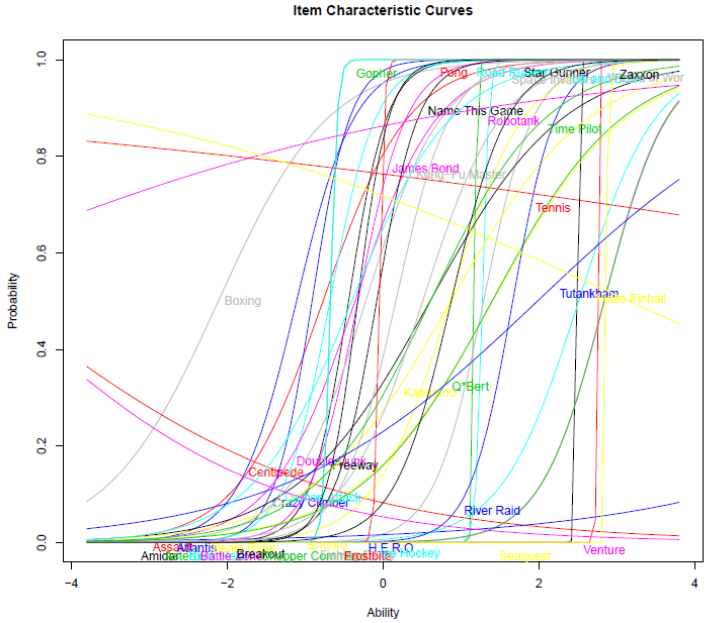
- How can we understand/improve items in a test?
  - Item Response Theory:
    - Logistic models: difficulty *b*, discrimination *a* and guessing *c*.

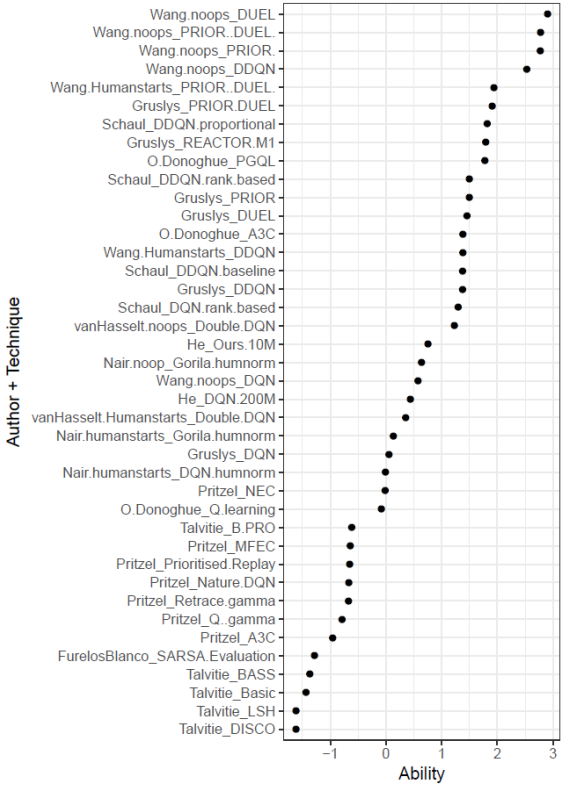


Proficiency (ability) as achievable difficulty

ITEM RESPONSE THEORY: APPLICATION TO ML/AI

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





Martínez-Plumed, F. et al. "Making sense of item response theory in machine learning", ECAI 2016, best paper award.

Martínez-Plumed, F., Hernández-Orallo, J. "AI results for the Atari 2600 games: difficulty and discrimination using IRT", EGPAI@IJCAI 2017.

Martínez-Plumed, F. "Item response theory in AI: Analysing machine learning classifiers at the instance level" Artificial Intelligence, 2019.

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

	μ_1	μ_2	μ_3	μ_4	μ_5
π_a	0.85	0.75	0.80	0.85	0.75
π_b	1.00	1.00	0.00	1.00	1.00

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

A Measure of Generality: It's all about difficulty

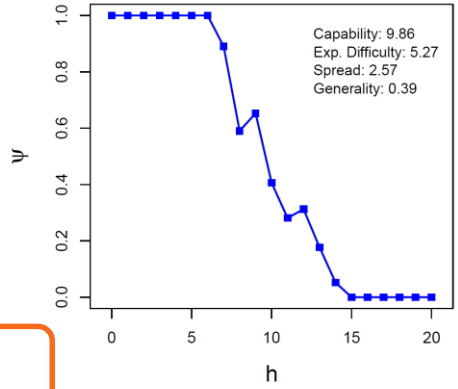
General intelligence must be seen as competence for a wide range of cognitive tasks up to a certain level of difficulty.

- Definition
  - Capability (Ψ), the area under the ACC: $\psi_j \triangleq \int_0^\infty \psi_j^{[h]}\, dh$
  - Expected difficulty given success:

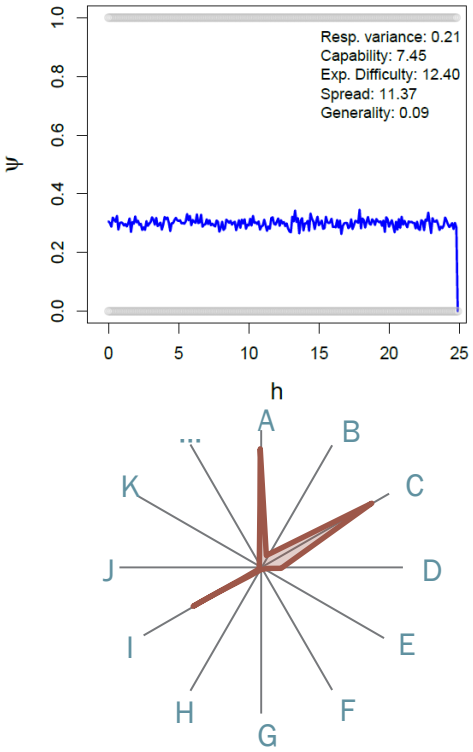
$$\mathbb{H}_j \triangleq \mathbb{E}_i[h|A_{i,j}=1] = \frac{m_j}{\psi_j} \qquad m_j \triangleq \int_0^\infty h \cdot \psi_j^{[h]}\, dh$$

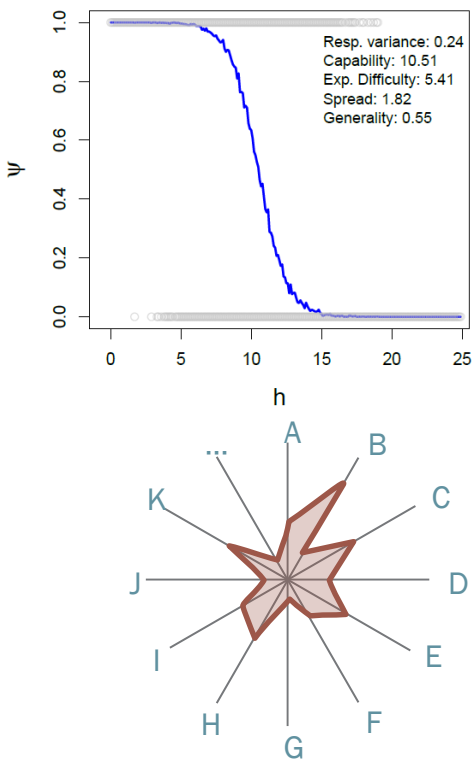
  - Spread: $z_j \triangleq \sqrt{(2\mathbb{H}_j - \psi_j)\cdot\psi_j} = \sqrt{2m_j - \psi_j^2}$
  - Generality: $\gamma_j \triangleq \frac{1}{z_j} = \frac{1}{\sqrt{2m_j - \psi_j^2}}$

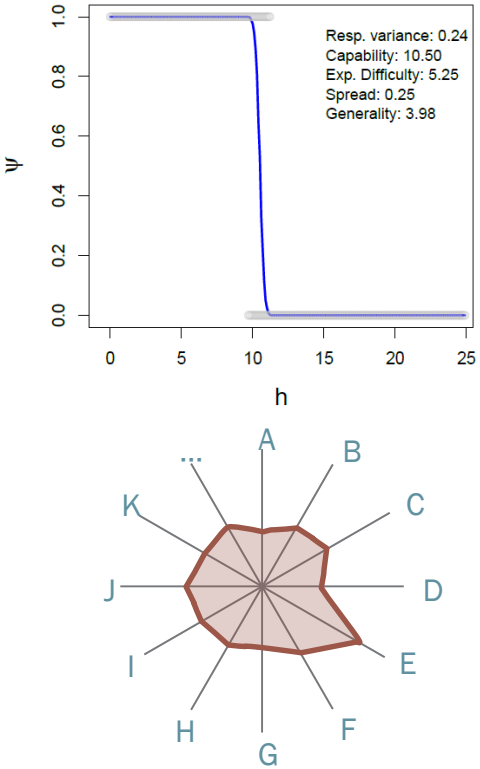
Non-populational!



A Measure of Generality: Some Agent Characteristic Curves





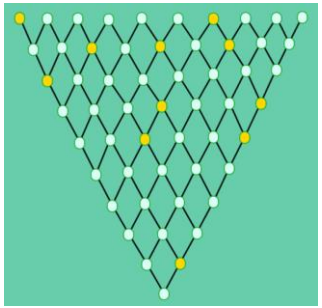


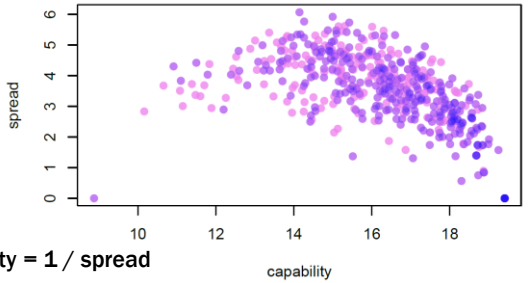
A Measure of Generality: Examples with Humans

- Capability and generality are observables, applied to individuals, no models.
- 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.



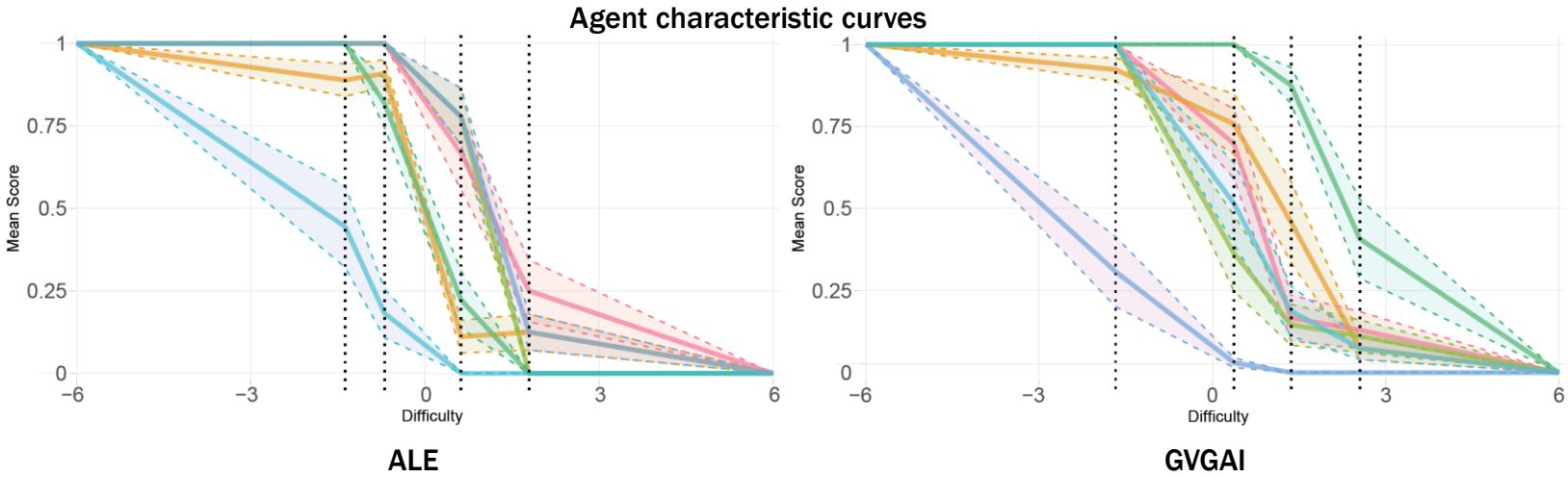


Generality = 1 / spread

RA Buckingham, A Elithorn, DN Lee, and WLB Nixon. A mathematical model of a perceptual maze test. Nature, 1963.
Davies A.D.M. and Davies, M.G. The difficulty and graded scoring of Elithorn's perceptual maze test. British Journal of Psychology, 1965

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 are systems more general?



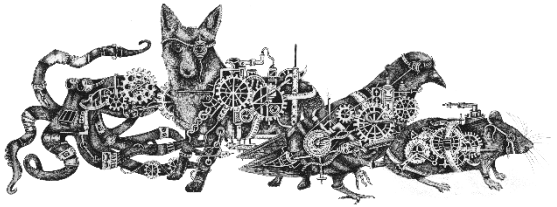
Martínez-Plumed, F., Hernández-Orallo, J. "Dual Indicators to Analyse AI Benchmarks: Difficulty, Discrimination, Ability and Generality", IEEE Transactions on Games 2018.

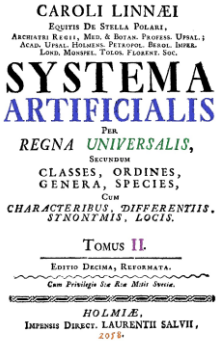
CONCLUSIONS

- 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, scaling, 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

- Paradigms of AGI and Their Associated Risks @ CSER:
  - How does generality affect AGI safety, together with capability and resources?
    https://www.cser.ac.uk/research/paradigms-AGI/
- The Atlas of Intelligence @ CFI:
  - Collection of maps comparing humans, non-human animals and AI systems.
- The AnimalAI Olympics @ CFI
  - http://animalaiolympics.com/





THANK YOU!