Integrating the Evaluation of Artificial and Natural Intelligence: *Are We Ready Yet?*

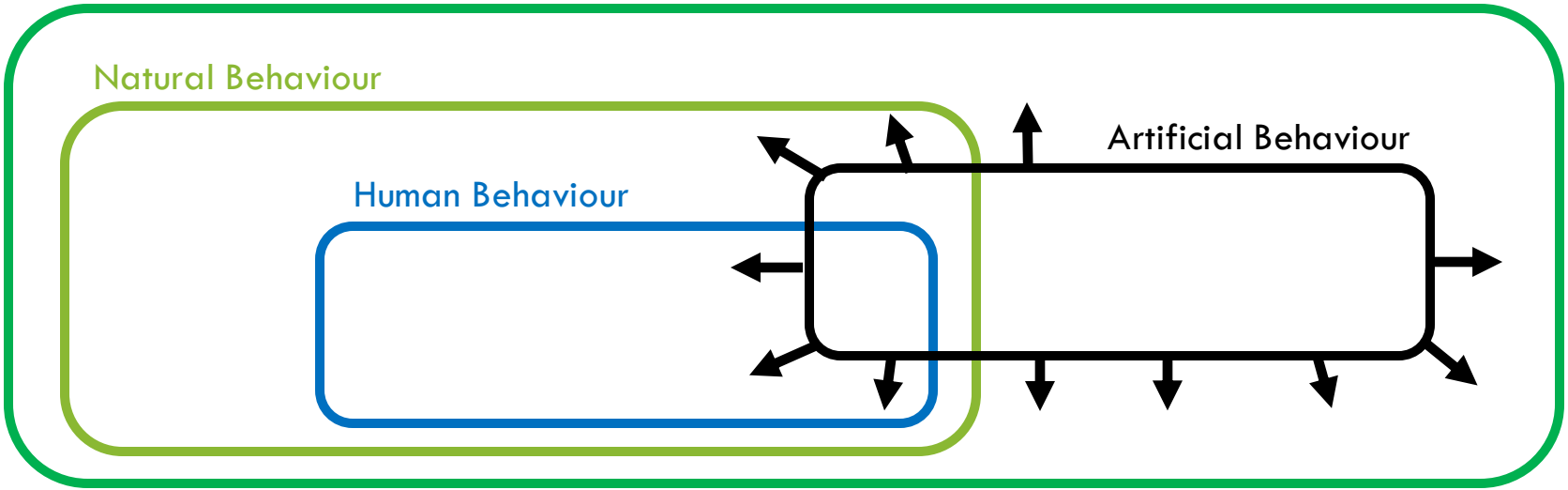
José Hernández-Orallo
Universitat Politècnica de València
Valencian Research Institute for Artificial Intelligence (VRAIN)
Valencian Graduate School and Research Network of AI
Leverhulme Centre for the Future of Intelligence
Centre for the Study of Existential Risk
jorallo@upv.es http://josephorallo.webs.upv.es/

Workshop on Understanding and Evaluating Big Models for Human Intelligence and Learning
April 11, 2023

A COPERNICAN REVOLUTION

- Where is **artificial intelligence** heading?

Sloman, A. "The structure of the space of possible minds " in The Mind and the Machine: philosophical aspects of Artificial Intelligence, Ed. S. Torrance, Ellis Horwood, 1984, pp 35-42.



MEASURING INTELLIGENCE

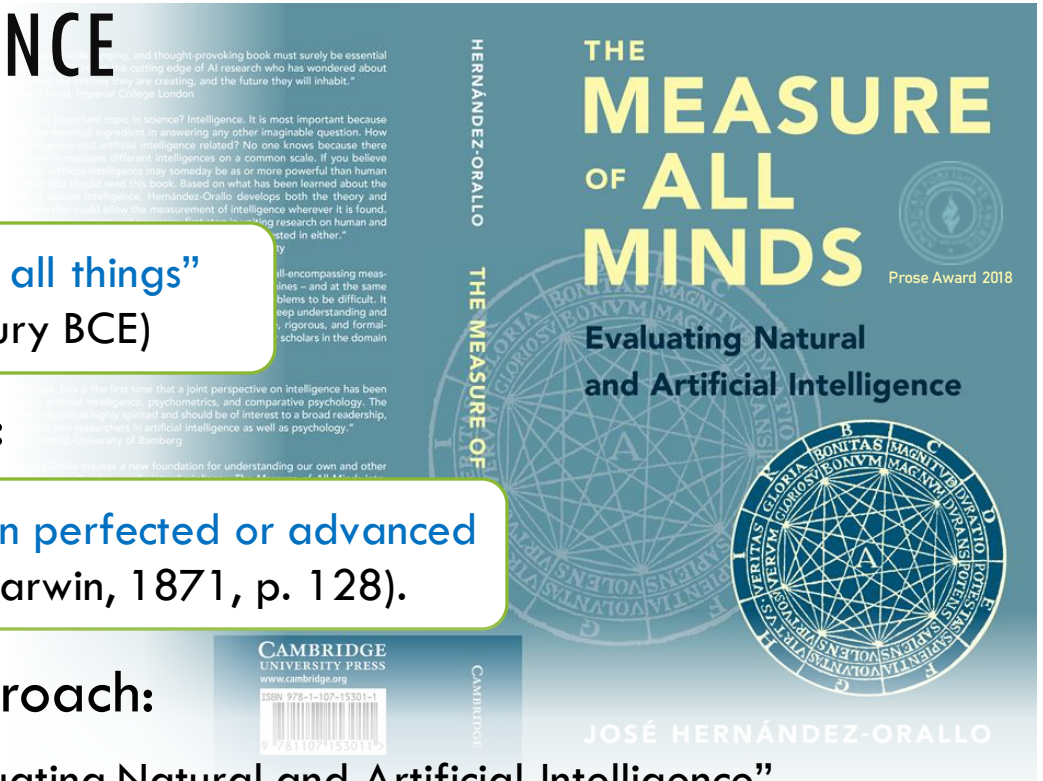
- From anthropocentrism:

> "Man is the measure of all things"
> (Protagoras, 5th century BCE)

- Or even from biocentrism:

> [intellectual faculties] "have been perfected or advanced through natural selection" (Darwin, 1871, p. 128).

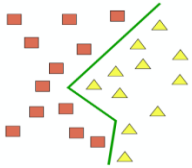
- To a more principled approach:
  - "The Measure of All Minds: Evaluating Natural and Artificial Intelligence", *Cambridge University Press*, 2017. http://www.allminds.org



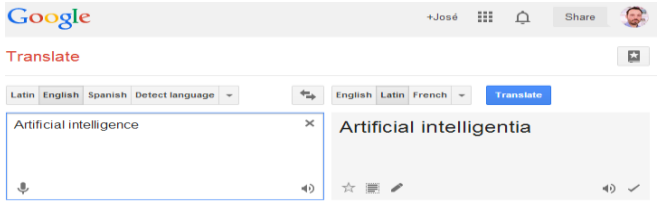
TASK-ORIENTED EVALUATION?

Specific (task-oriented) AI systems

PR: computer vision, speech recognition, etc.



Prediction and estimation

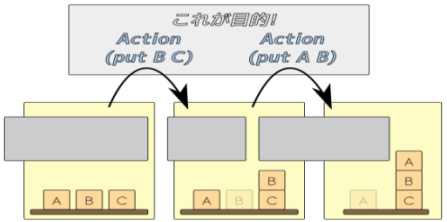


Machine translation, information retrieval, summarisation

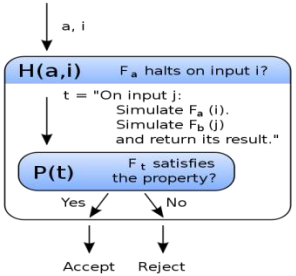
Robotic navigation

Knowledge-based assistants

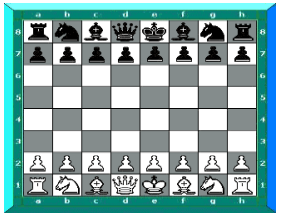
Driverless vehicles



Planning and scheduling



Automated deduction



Game playing

All images from wikicommons

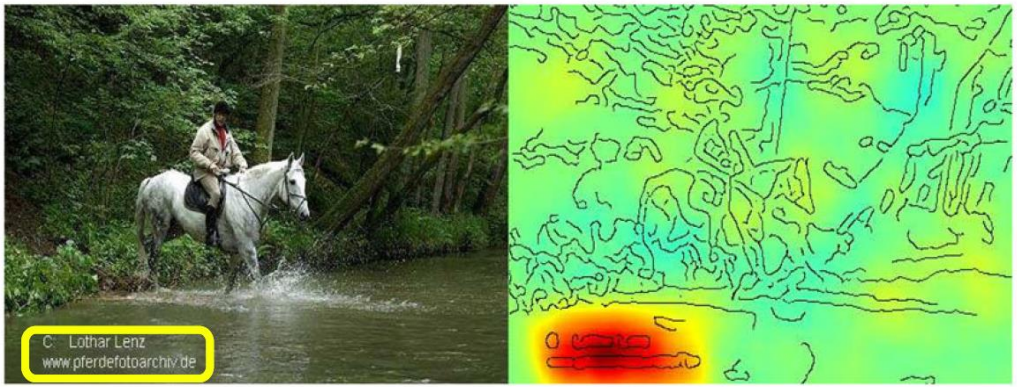
PERFORMANCE ON THE TASK WITHOUT THE CAPABILITY

- Benchmarks collect **particular task distributions** – AI overfits:
  - Adversarial examples
  - Clever Hans phenomenon:

Hernández-Orallo, J. et al. "A New AI Evaluation Cosmos: Ready to Play the Game?" AI Magazine 38 (3), 2017.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1-8.

Horse-picture from Pascal VOC data set

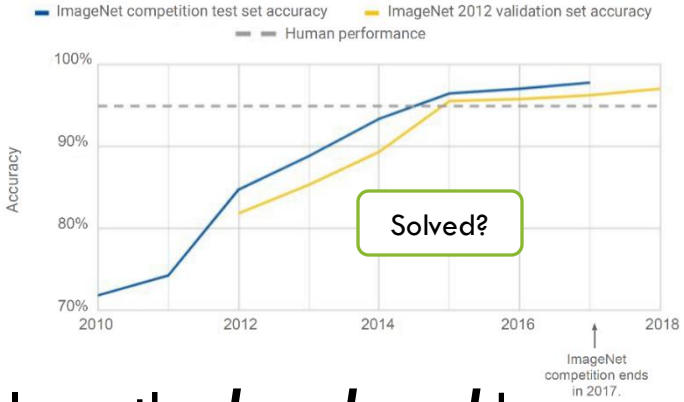


Hernández-Orallo, J. (2019). Gazing into Clever Hans machines. *Nature Machine Intelligence*, *1*(4), 172-173.

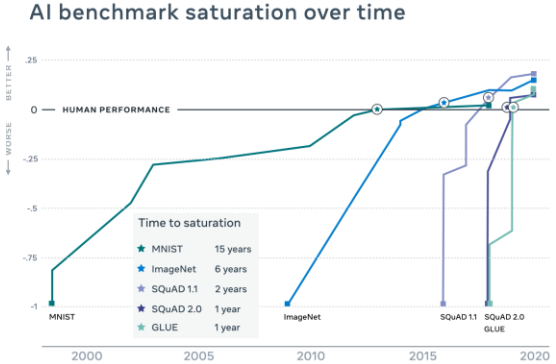
F Martínez-Plumed, P. Barredo, S. Ó hÉigeartaigh, J. Hernández-Orallo (2021) "Research community dynamics behind popular AI Benchmarks" *Nature Mach Int.* https://rdcu.be/ckK8X

NOT ONLY OVERFITTING, BUT A SCALE PROBLEM

- AI results become superhuman, but AI doesn't have the capability.



Hernandez-Orallo, J. "AI Evaluation: On Broken Yardsticks and Measurement Scales", MetaEval@AAAI2020.



Give me the data (distribution) and I will ace the test in a year!

From: https://ai.facebook.com/blog/dynabench-rethinking-ai-benchmarking

- Replace the ***benchmark***!

'challenge-solve-and-replace' (Schlangen, 2019), or a 'dataset-solve-and-patch' (Zellers et al., 2019) dynamics.

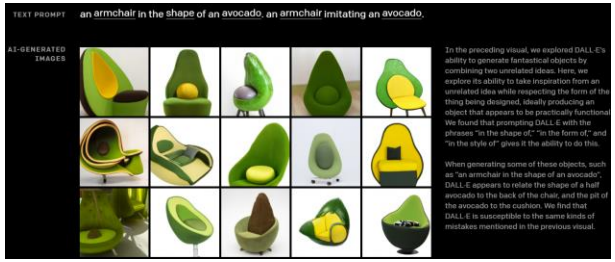
CIFAR10 → CIFAR100,
SQuAD1.1 → SQuAD2.0,
GLUE → SUPERGLUE,
Starcraft → Starcraft II

Date	Model	EM	F1
	Humans	86.83	89.45
Dec 13, 2018	BERT finetune	83.54	86.10
April 06, 2020	SA-Net on Albert	**90.72**	**93.01**

CAPABILITY-ORIENTED EVALUATION: WHY?

general-purpose systems and cognitive components

Cognitive robots

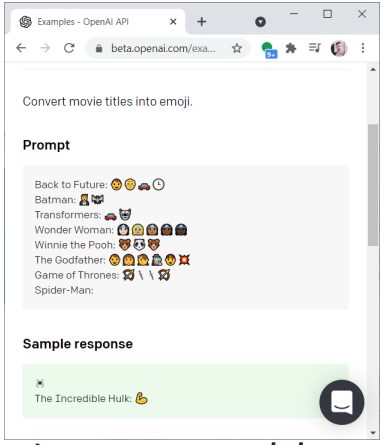


Designers, creators

Smart environments

Agents, avatars, chatbots

Personal assistants

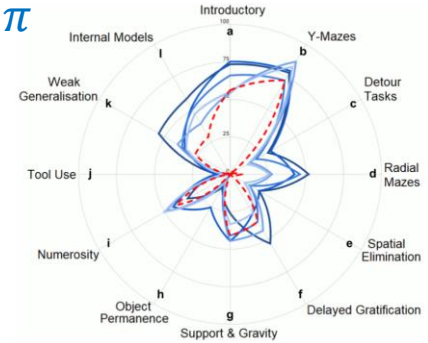


Language models

CAPABILITY-ORIENTED EVALUATION: WHAT?

- Infers **capabilities** for system π and **demands** for problem μ.

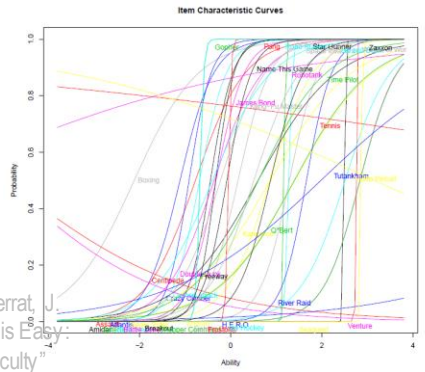
π



Crosby, M., Beyret, B., Shanahan, M., Hernández-Orallo, J., Cheke, L., & Halina, M. (2020). The animal-AI testbed and competition. In NeurIPS 2019 Competition and Demonstration Track (pp. 164-176). PMLR.

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

μ



F. Martínez-Plumed, D. Castellano, C. Monserrat, J. Hernández-Orallo (2021) "When AI Difficulty is Easy: The Explanatory Power of Predicting IRT Difficulty"

F Martinez-Plumed, J Hernandez-Orallo (2020) "Dual indicators to analyse AI benchmarks: Difficulty, discrimination, ability and generality" *IEEE Transactions on Games. 12(2), 121 - 131*

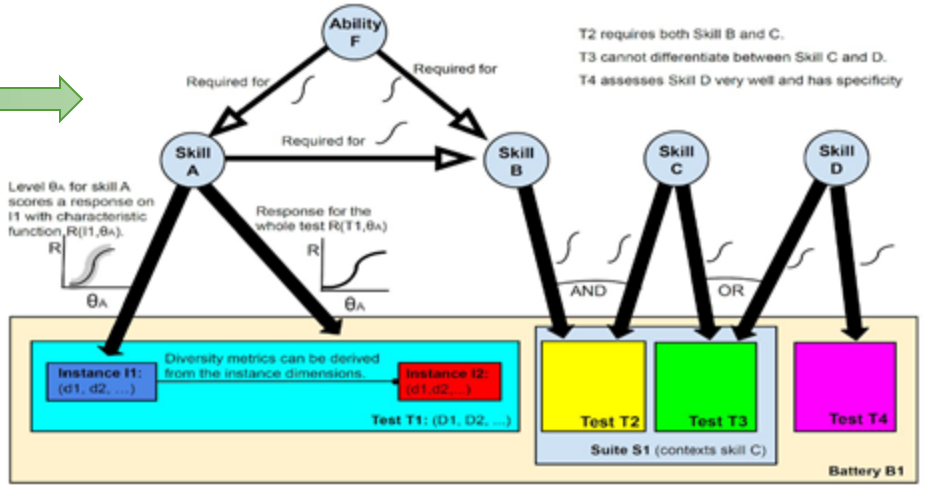
Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., & Hernández-Orallo, J. (2019). Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271, 18-42.

- Key ideas:
  - Instance difficulties (demands) become dual to capabilities (à la IRT).

No normative population: IRT? FA? SEM?

RECOG-AI : MEASUREMENT LAYOUTS

- Robust Evaluation of Cognitive Capabilities and Generality in AI
  - 2021-2024 project:
    - http://lcfi.ac.uk/projects/kinds-of-intelligence/recog-ai/
    - Run at the Centre for the Future of Intelligence, Cambridge, UK.
  - Measurement Layouts:
    -
  - Generality:
    - In RL settings for basic navigation skills
    - With language or multimodal models



RECOG-AI : SPACES AND FEATURES

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211-252.

- Original feature space:
  - observable by the system. Usually abstracted into latent features.

128x128 RGB pixels

- Surface feature space:
  - sometimes observable. A general system should be invariant to these.

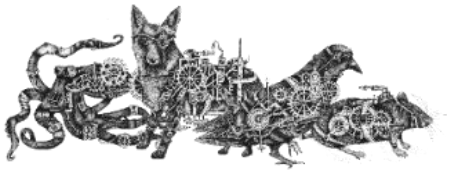
Symmetry

Irrelevant elements

- Cognitive (construct) space:
  - usually non-observable. Performance should correlate with them:
    - high-capabilities agents should imply success for problems with lower difficulty levels in these capabilities.

WP: Assimilate existing benchmarks to these layouts

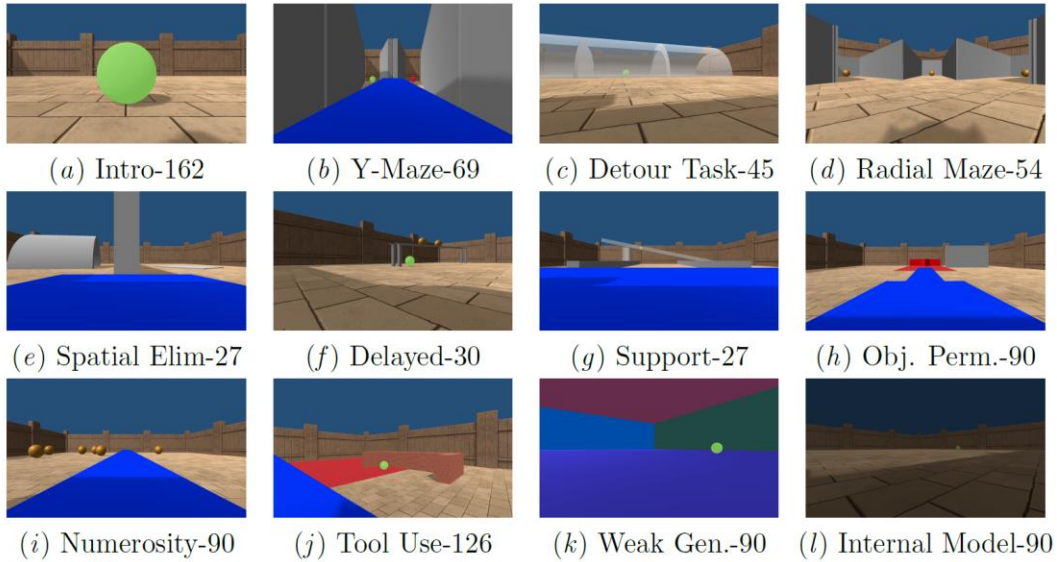
RECOG-AI: ANIMAL AI SANDBOX



Goal: confront AI systems with tasks commonly used in comparative cognition, especially in animal cognition

http://animalai.org/

Crosby, M., Beyret, B., Shanahan, M., Hernández-Orallo, J., Cheke, L., & Halina, M. (2020, August). The animal-ai testbed and competition. In *NeurIPS 2019 Competition and Demonstration Track* (pp. 164-176). PMLR.

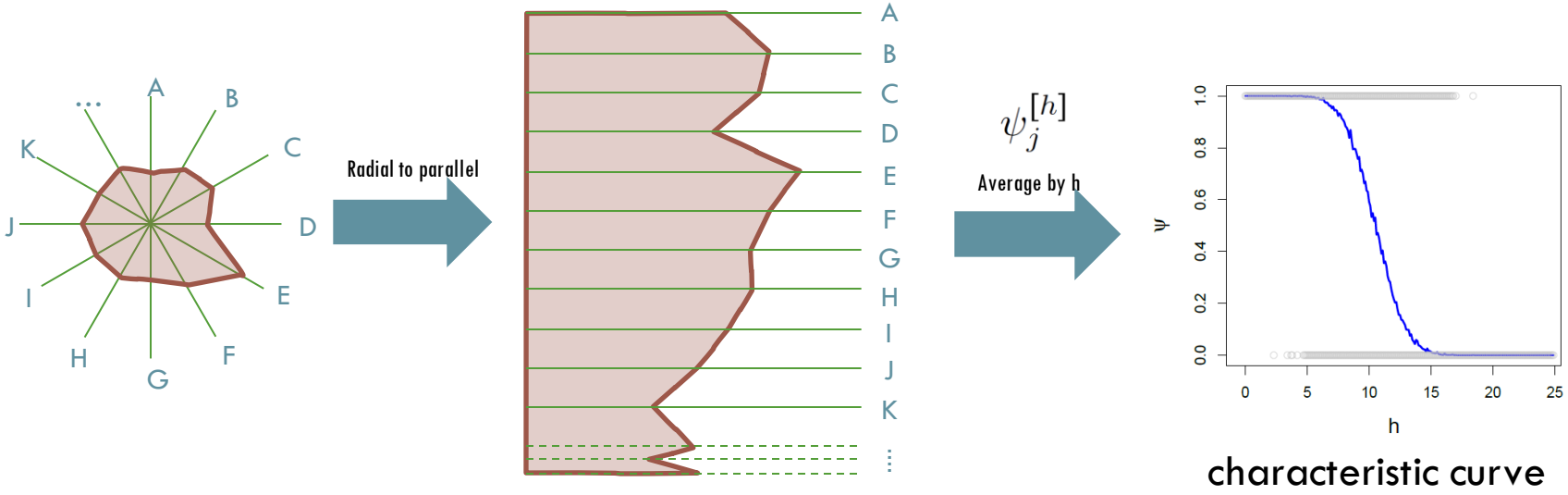


Example problems from each of the 12 task types. (a) contains familiarisation tasks. (b-j) have direct links to animal tests. (k,l) are AI-specific. Number of problems per category shown in captions (900 total).

GENERALITY

Hernandez-Orallo, J.; Loe, B.S.; Cheke, L.; Martínez-Plumed, F., O h´Eigeartaigh, S. "General intelligence disentangled via a generality metric for natural and artificial intelligence", Nature Sci Rep 2021

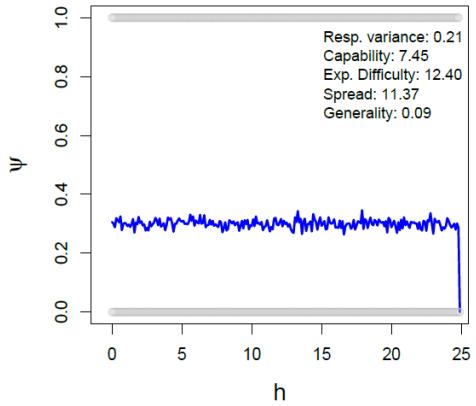
- *Systematic performance for a range of tasks up* ***to a level of difficulty.***
  - Radial capability profiles can be aggregated into one **characteristic curve:**
    - All dimensions A, B, C, … are made commensurate by policy difficulty:

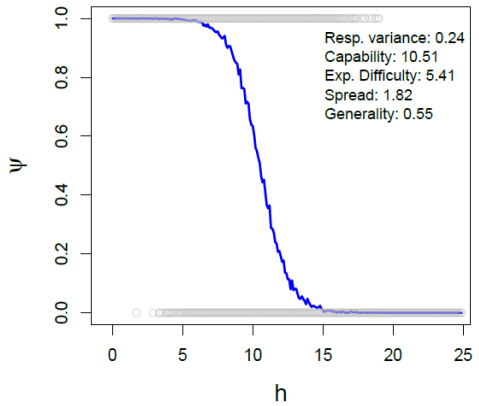


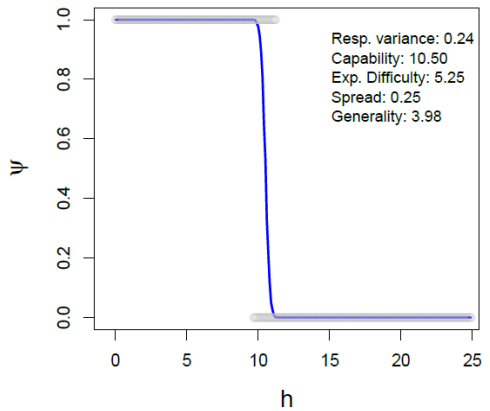
characteristic curve

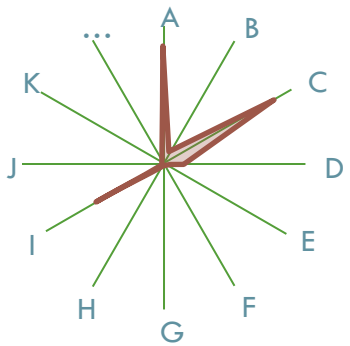
A MEASURE OF GENERALITY

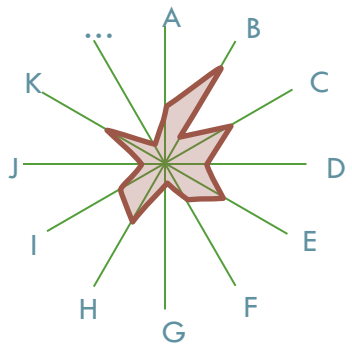
Hernandez-Orallo, J.; Loe, B.S.; Cheke, L.; Martínez-Plumed, F., O h´Eigeartaigh, S. "General intelligence disentangled via a generality metric for natural and artificial intelligence", Nature Sci Rep 2021

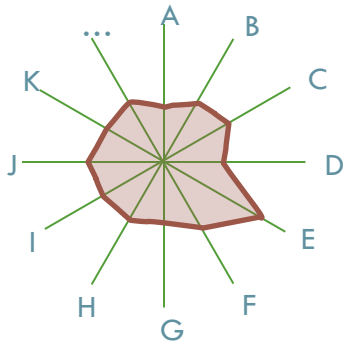










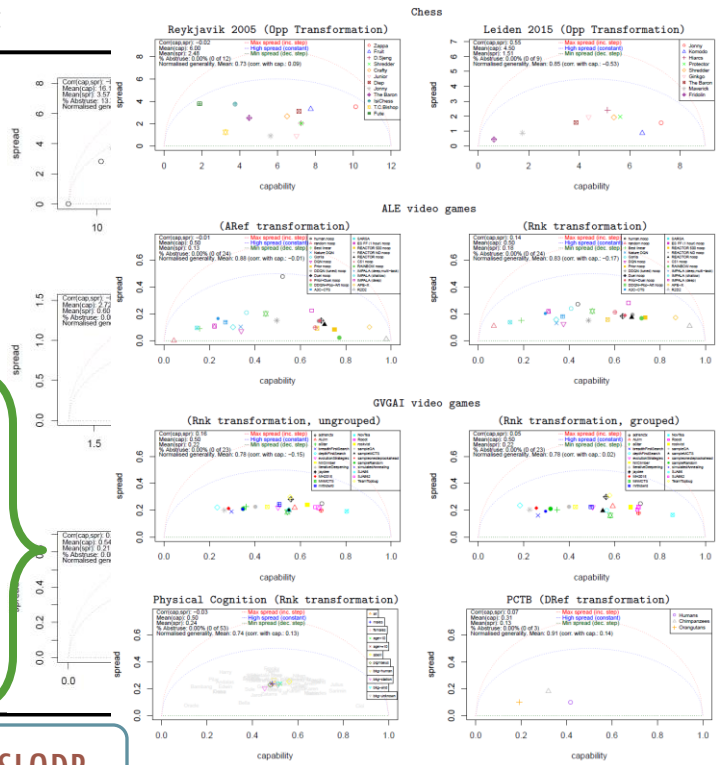


GENERALITY: SLODR?

Hernandez-Orallo, J.; Loe, B.S.; Cheke, L.; Martínez-Plumed, F., O h´Eigeartaigh, S. "General intelligence disentangled via a generality metric for natural and artificial intelligence", Nature Sci Rep 2021

- Results for many scenarios and subjects:

Scenario	Subjects
Elithorn's mazes	Humans
Letter series	Humans and Machines
Object recognition	Humans, Macaques and Machines
Odour span task	Rats
Iris (KDM)	Machines
Iris (TD_U)	Machines
Chess (Reykjavik)	Machines
Chess (Leiden)	Machines
ALE video games	Machines (+ Human reference)
ALE video games	Machines and Human
GVGAI games (agg.)	Machines
Physical Cognition	Orangutans
PCTB	Humans, Chimpanzees and Orangutans

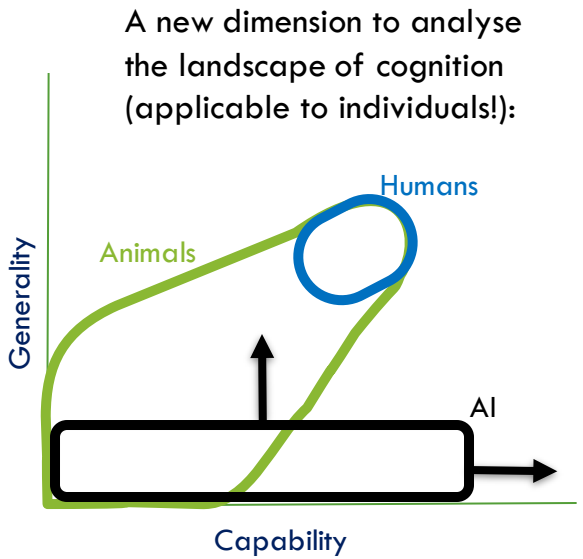


We found no evidence of lower generality for high-ability groups:

No SLODR.

DISENTANGLING GENERALITY AND CAPABILITY

- Generality as a measure that is different from capability:



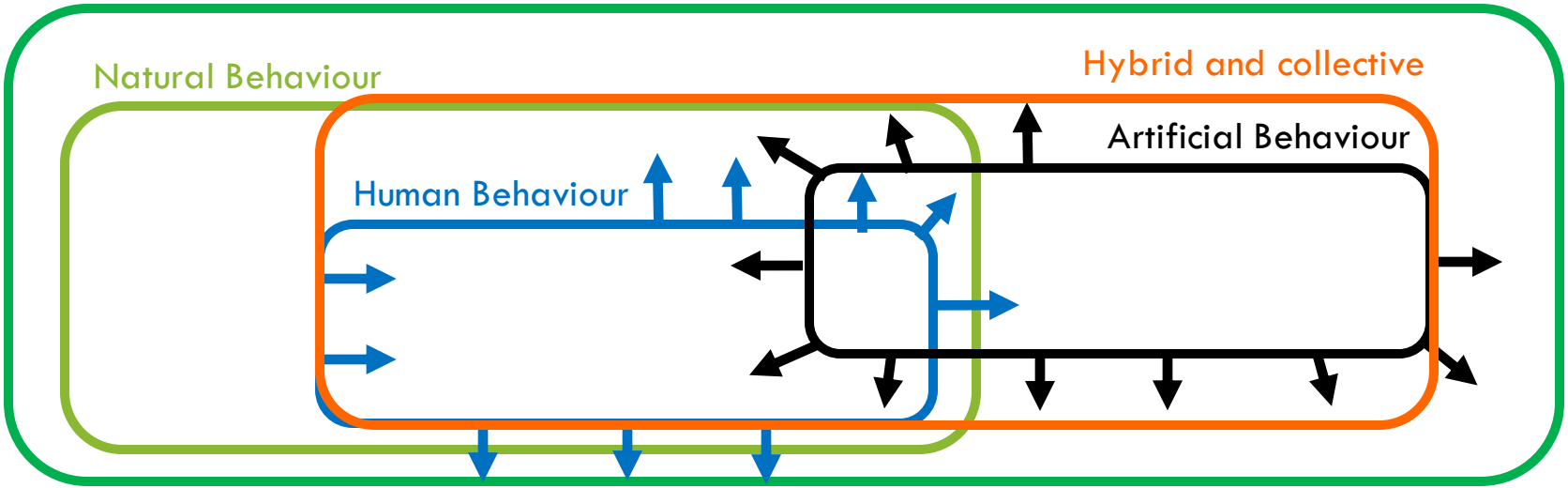
All this requires the **identification of difficulty** based on perceptual/cognitive elements or required resources!

HUMAN COGNITION IS CHANGING TOO

- Where are all kinds of **intelligence** heading?
  - Human, artificial, hybrid, collective, …

Can we measure this?

Sloman, A. "The structure of the space of possible minds " in The Mind and the Machine: philosophical aspects of Artificial Intelligence, Ed. S. Torrance, Ellis Horwood, 1984, pp 35-42.



GENERAL-PURPOSE AI: IS IT HERE?

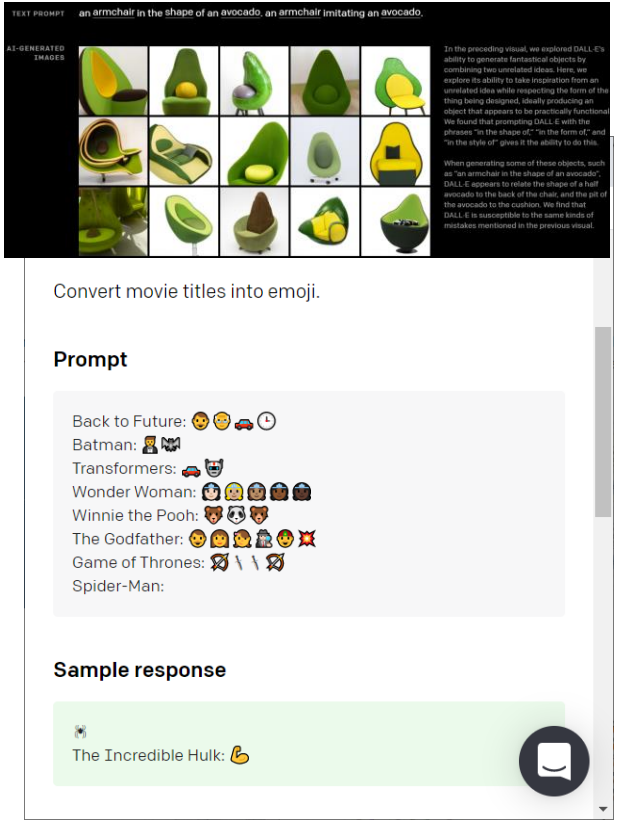
- Massive Multimodal Models:

  (now being referred to as Foundation Models)

  - They compress human culture
  - First evaluated by perplexity (compressors)
  - Surprising zero-shot and few-shot inference:
    - Off-the-shelf systems that are really *general*?
    - Their quality depends on their "prompts"

Bommasani et al. "On the Opportunities and Risks of Foundation Models." *arXiv preprint arXiv:2108.07258* (2021).

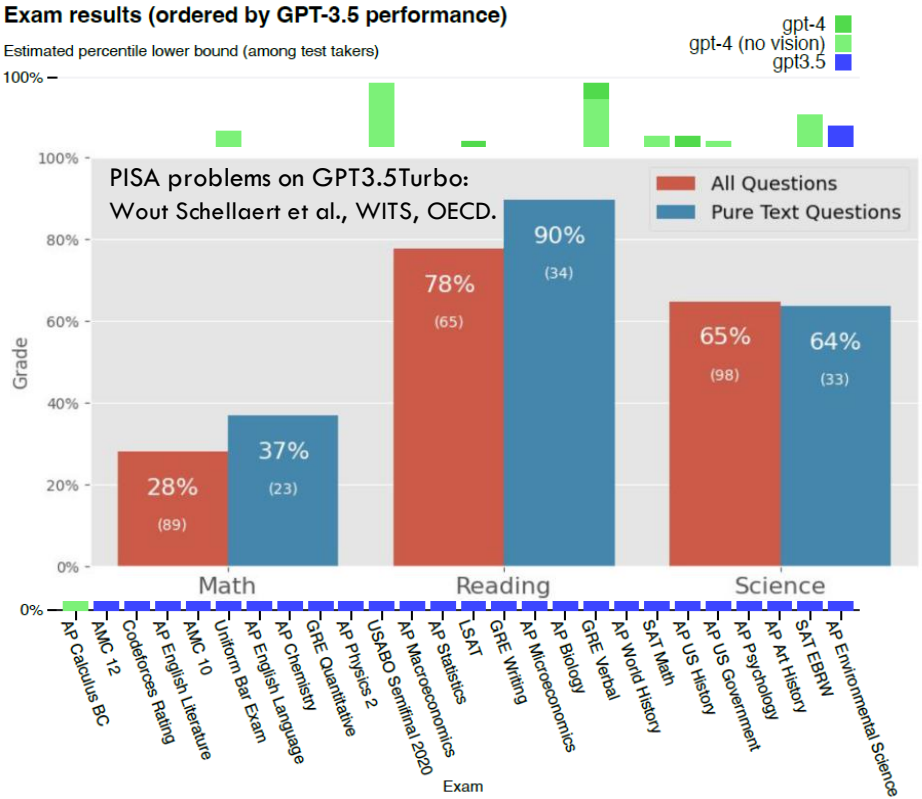
How can we evaluate these systems?



GENERAL-PURPOSE AI. HOW NOT TO MEASURE IT?

- Using collections of AI benchmarks?
- Using psychometric tests?
- Using achievement assessments?
- Using non-contaminated assessments?

Misconceptions that should have been left behind years ago!

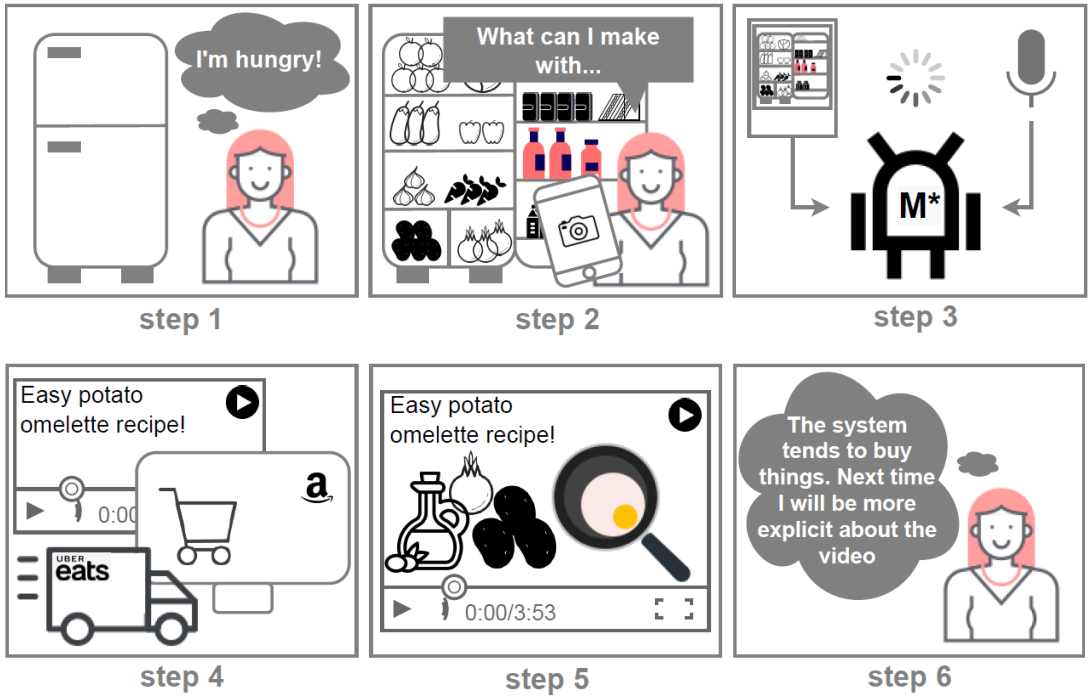


GENERAL-PURPOSE AI. HOW TO MEASURE IT?

P. Moreno, B.S. Loe, J. Burden, S. Ó hÉigeartaigh, J. Hernandez-Orallo "How General-Purpose Is a Language Model? Usefulness and Safety with Human Prompters in the Wild", AAAI2022.

- They are assistants:

What matters for AI evaluation, human skills, etc., is the **hybrid human-AI system**!



Schellaert, Plumed, Vold, Burden, Casares, Loe, Reichart, OhEigeartaigh, Korhonen, Orallo "Your Prompt is My Command: Assessing the Human-Centred Generality of Multi-Modal Models". JAIR, 2023,

READY TO GET READY!

- Big models have brought general-purpose AI
  - Performance-based evaluation must leave way to **capability-oriented evaluation**
  - **Psychometrics, psychology and cognitive science** finally vindicated!
  - The techniques require **significant adaptation and reinterpretation** (no AI populations!)
- Also, this new AI is mostly assistive – hybrid – augmentative
  - What matters is the **symbiosis human-AI**
  - Human **skills are also changing** quickly
  - Education must embrace **AI-extended students and professionals**

> It may someday happen . . . that the fields of artificial and human intelligence will grow closer together, each learning from the other.
>
> *– Douglas K. Detterman, A Challenge to Watson (2011)*

Thank you!

José Hernández Orallo
http://josephorallo.webs.upv.es/
jorallo@upv.es

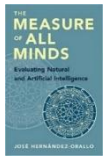
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

VRAIN

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

Other Talks (http://josephorallo.webs.upv.es/)

- Diversity Unites Intelligence: Measuring Generality
- Measuring A(G)I Right: Some Theoretical and Practical Considerations
- Natural and Artificial Intelligence: Measures, Maps and Taxonomies



Book (http://allminds.org):

- "The Measure of All Minds: Evaluating Natural and Artificial Intelligence", Cambridge U.P. http://allminds.org

The AI Collaboratory: https://ai-collaboratory.jrc.ec.europa.eu/ (old: http://dmip.webs.upv.es/AICollaboratory/)

[A]I collaboratory

- Part of the European Commission's AI watch: https://ec.europa.eu/knowledge4policy/ai-watch_en
- Technology Readiness Levels: https://data.europa.eu/doi/10.2760/495140
- Measuring the Occupational Impact of AI: https://jair.org/index.php/jair/article/view/12647

European Commission

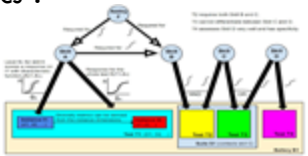
OECD's AI and the Future of Skills Project:

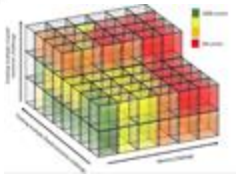
- https://www.oecd.org/education/ceri/Future-of-Skills-Overview.pdf, https://doi.org/10.1787/5ee71f34-en.

OECD

DARPA RECoG-AI Project: http://lcfi.ac.uk/projects/kinds-of-intelligence/recog-ai/

- Part of the Kinds of Intelligence Programme at the CFI in Cambridge
  - http://lcfi.ac.uk/projects/kinds-of-intelligence
- IJCAI2022 Workshop "AI Evaluation Beyond Metrics":
  - https://sites.google.com/view/ebem2022





WE'RE HIRING!