ASSESSING AI CAPABILITIES: TOO IMPORTANT TO BE FOOLED

José Hernández-Orallo (jorallo@upv.es, josephorallo.webs.upv.es)
Valencian Research Institute for Artificial Intelligence (vrAIn) (vrain.upv.es)
Universitat Politècnica de València, València (www.upv.es)
Leverhulme Centre for the Future of Intelligence, Cambridge (lcfi.ac.uk)

VRAIN

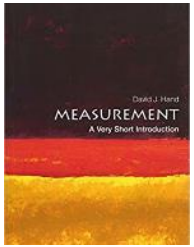
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

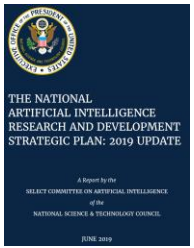
OECD International Conference on AI in Work, Innovation, Productivity and Skills (AI-WIPS): 1-5 February 2021
Panel on exploring assessments of AI capabilities: 4 February 2021

https://oecd-events.org/ai-wips/

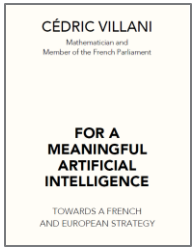
ASSESSING AI CAPABILITIES. IT IS VERY IMPORTANT!



"Greatest accuracy, at the frontiers of science, requires greatest effort, and probably the most expensive or complicated of measurement instruments"
David Hand, "Measurement: A Very Short Introduction", OUP, (2004)



"Measure and Evaluate AI Technologies through Standards and Benchmarks"
Strategy 6 in U.S. National
AI Research and Development Strategic Plan: (2019)



"Public authorities must act in order to develop and implement standards, tests and measurement methods [for] AI technology"
Villani report (French AI Strategy): (2018)

WHY IS IT REALLY SO IMPORTANT?

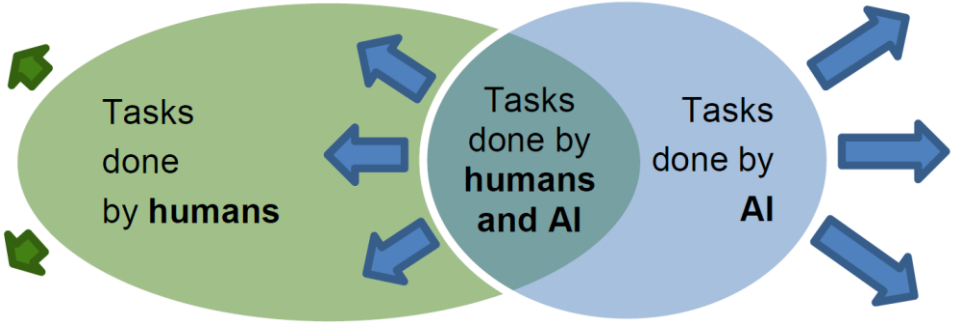
- AI is not just another emergent technology. It is about **intelligence**:
  - What has made us "understand" and attempt to "control" nature.
- It is going to **transform** every sphere of society radically:
  - From education to the workplace.
- But:

> The way and the extent the world is affected by AI depends on **what and how much AI can do!**

- Key question for policies, regulations, investments, opportunities, risks…
  - We need to evaluate what a system (AI or hybrid) **can** and **cannot** do.

WHAT IS AI EVALUATION?

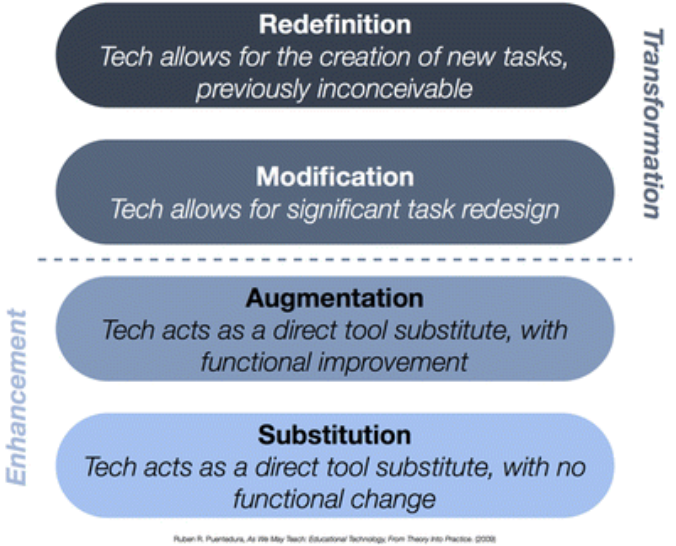
- Evaluating AI on tasks humans do today?
  - The "substitution narrative" is too narrow!



  - It is not just what *tasks AI systems can do*, but *what skills AI systems can acquire by themselves*

"The right kind of AI"?

Specific vs. General AI Systems



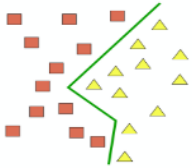
Puentedura, R. (2014b). Learning, technology, and the SAMR model: Goals, processes, and practice [Blog post]. http://www.hippasus.com/rpweblog/archives/2014/06/29/LearningTechnologySAMRModel.pdf. Hamilton, E.R., Rosenberg, J.M. & Akcaoglu, M. The Substitution Augmentation Modification Redefinition (SAMR) Model: a Critical Review and Suggestions for its Use. TechTrends 60, 433–441 (2016). https://doi.org/10.1007/s11528-016-0091-y

Acemoglu, D, and Restrepo, P. "The wrong kind of AI? Artificial intelligence and the future of labour demand." *Cambridge Journal of Regions, Economy and Society* 13.1 (2020): 25-35.

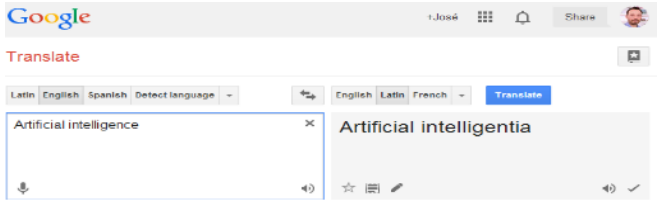
TASK-ORIENTED EVALUATION?

- Specific (task-oriented) AI systems

PR: computer vision, speech recognition, etc.



Prediction and estimation

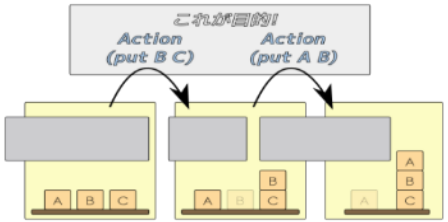


Machine translation, information retrieval, summarisation

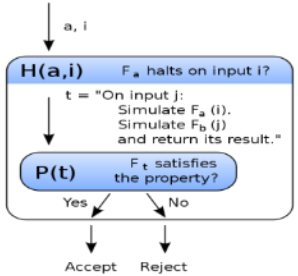
Robotic navigation

Knowledge-based assistants

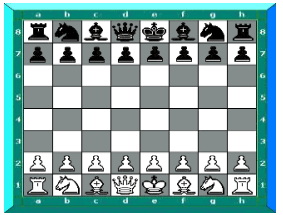
Driverless vehicles



Planning and scheduling



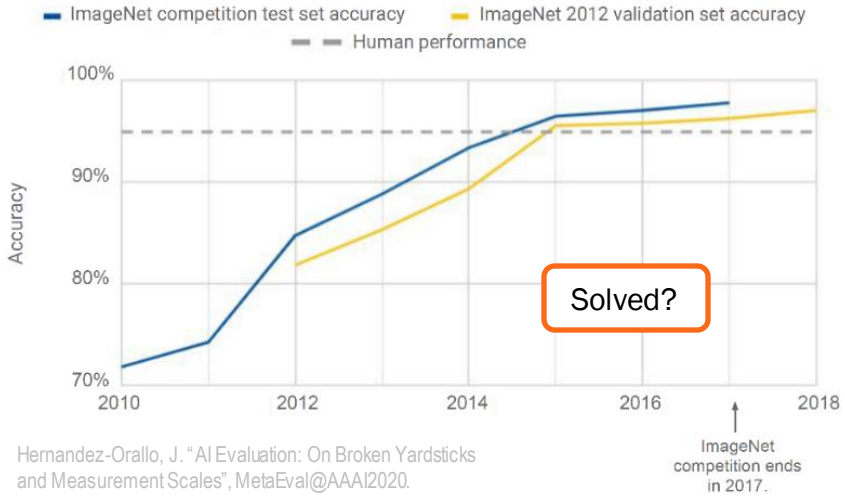
Automated deduction



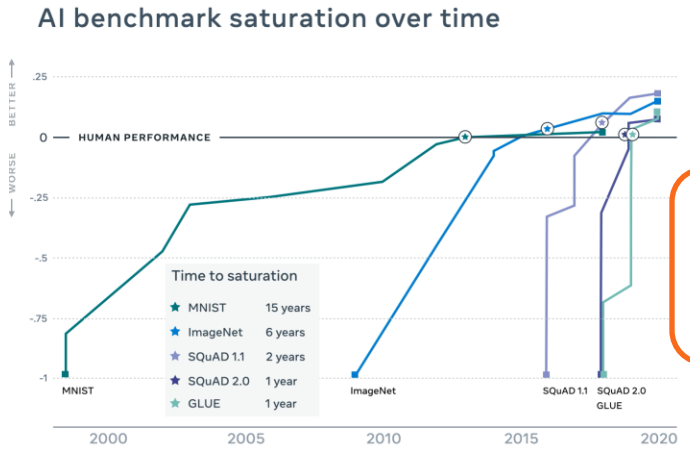
Game playing

All images from wikicommons

TASK-ORIENTED EVALUATION: TEACHING TO THE BENCHMARK!



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



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

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

MISLEADING!!!

Can we anticipate how these systems will behave for new problems or other distributions of the same problem?

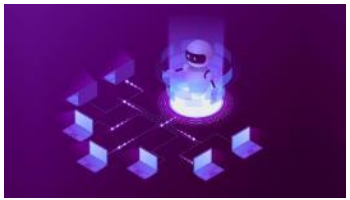
CAPABILITY-ORIENTED EVALUATION?

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

Cognitive robots

Agents, avatars, chatbots

Pets, animats and other artificial companions



Internet bots, Smartbots, Security bots…

Smart environments

Personal assistants

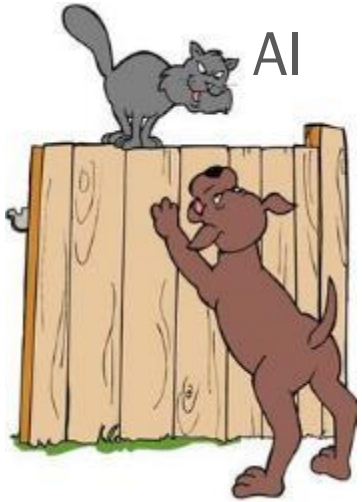
CAPABILITY-ORIENTED EVALUATION: MANY APPROACHES!

- "AI-completeness" benchmarks:
  - Science exams, commonsense reasoning
- The "Mythical" Turing Test:
  - And a myriad variants….
- New evaluation platforms:
  - Videogames, naïve physics, etc.
- Psychometric tests:
  - IQ tests, developmental tests, …
- Comparative cognition (animal) tests:
  - Morgan's canon?

TOO IMPORTANT TO BE FOOLED!

- A call for action: new AI evaluation initiatives!
  - More principled, more robust.
  - More interdisciplinary: AI, psychology, cognitive/social sciences, …
  - More independent and international (e.g., OECD, JRC@EC, academia)
  - More long-term and integrated…

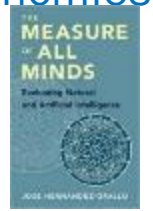
A challenge for the whole century!
We need to start now, as AI evaluation is lagging behind AI!



THANKS!

OTHER SOURCES AND INITIATIVES:

- 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 University Press 2017



- The AI Collaboratory: http://aicollaboratory.org/
  - Part of the European Commission's AI watch:
    - https://ec.europa.eu/knowledge4policy/ai-watch_en

[A]I collaboratory

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