ASSESSING AI CAPABILITIES:
TOO IMPORTANT TO BE FOOLED

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

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”

“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”
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

- **Redefinition**
  - Tech allows for the creation of new tasks, previously inconceivable

- **Modification**
  - Tech allows for significant task redesign

- **Augmentation**
  - Tech acts as a direct tool substitute, with functional improvement

- **Substitution**
  - Tech acts as a direct tool substitute, with no functional change

---

Specific vs. General AI Systems


**Task-oriented evaluation?**

- **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.*

*All images from wikicommons*
TASK-ORIENTED EVALUATION: TEACHING TO THE BENCHMARK!

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

MISLEADING!!!

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


Give me the data (distribution) and I will ace the test in a year!
CAPABILITY-ORIENTED EVALUATION?

How to evaluate general-purpose systems and cognitive components?

- Cognitive robots
- Pets, animats and other artificial companions
- Agents, avatars, chatbots
- Smart environments
- Internet bots, Smartbots, Security bots...
- 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/](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](http://allminds.org)):
  - The Measure of All Minds: Evaluating Natural and Artificial Intelligence, Cambridge University Press 2017

- The AI Collaboratory: [http://aicollaboratory.org/](http://aicollaboratory.org/)
  - Part of the European Commission’s AI watch:

- Other Events:
  - epAI (Evaluating progress in AI, at ECAI, September 2020)