IDENTIFYING ARTIFICIAL INTELLIGENCE CAPABILITIES: WHAT AND HOW TO TEST ?

José Hernández-Orallo (jorallo@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

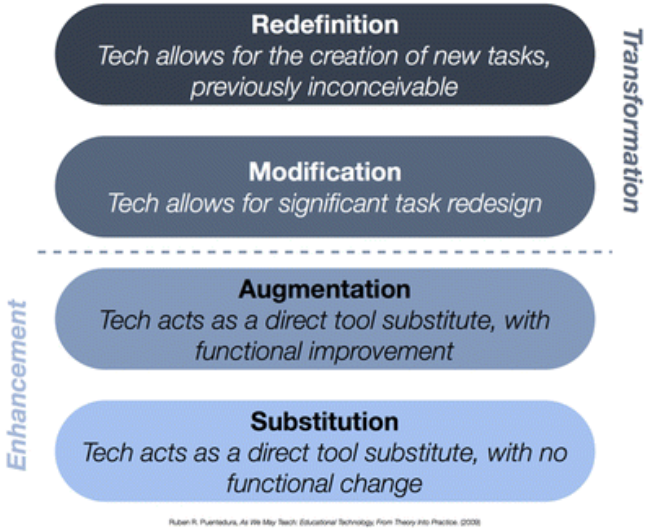
OECD Expert Meeting on Skills and Tests for Assessing AI and Robotics, 5-6 October 2020

Skills Are Changing

- Education must anticipate future societal and technological changes.

> Most (if not all) cognitive tasks human do will be done by AI in the future

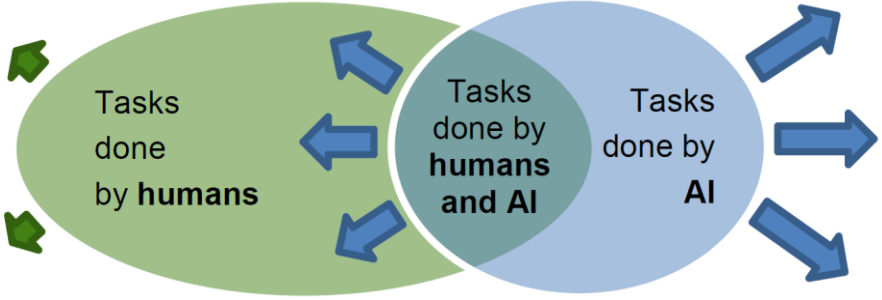
- Automation narratives about technology:
  - Replacing humans: “occupations replaced by robots”
  - Displacing humans: *fauxtomation*, human computation
  - Extending humans: AI extenders.



Puentedura, R. (2014b). Learning, technology, and the SAMR model: Goals, processes, and practice [Blog post]. http://www.hippasus.com/rrpweblog/archives/2014/06/29/LearningTechnologySAMRModel.pdf.

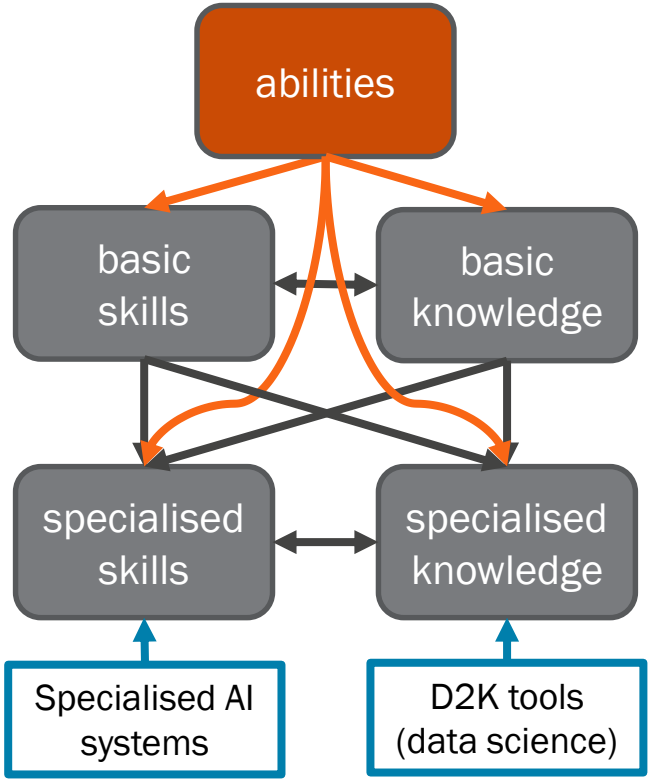
Hamilton, E.R., Rosenberg, J.M. & Akcaoglu, M. The Substitution Augmentation Modification Redefinition (SAMR) Model: a Critical Review and Suggestions for its Use. TechTrends 60, 433–441 (2016). https://doi.org/10.1007/s11528-016-0091-y

Skills Are Changing



- For a fast-changing situation, humans and AI should limit task/skill specialisation and aim at general abilities to acquire new skills.

More focus on abilities (and basic skills) rather than specialised skills and knowledge



IDENTIFYING CAPABILITIES: TAXONOMIES

Humans

- Many taxonomies of skills in occupational categories (O*NET-SOC, ISCO, ESCO, …)
  - By sectors (e.g., “armed forces”), by rank (e.g., “managers”) or generic (e.g., “professionals”).
- Cognitive abilities in human intelligence models and psychometrics.
  - E.g., Cattell-Horn-Carroll taxonomy.
- Developmental perspective
  - Skills develop over some other skills and abilities: sensorimotor, preoperational, concrete-operational, and formal-operational.

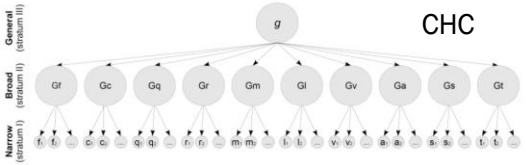
ISCO

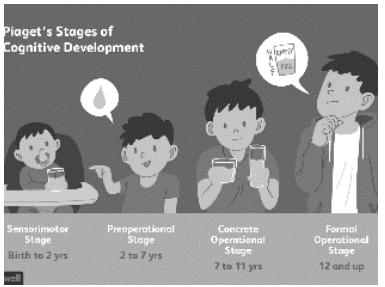
Code	Category
1	Managers
2	Professionals
3	Technicians and associate professionals
4	Clerical support workers
5	Service and sales workers
6	Skilled agricultural, forestry and fishery workers
7	Craft and related trades workers
8	Plant and machine operators, and assemblers
9	Elementary occupations
0	Armed forces occupations

ESCO

Code	Category
S1	Communication, collaboration and creativity
S2	Information skills
S3	Assisting and caring
S4	Management skills
S5	Working with computers
S6	Handling and moving
S7	Constructing
S8	Working with machinery and specialised equipment

CHC

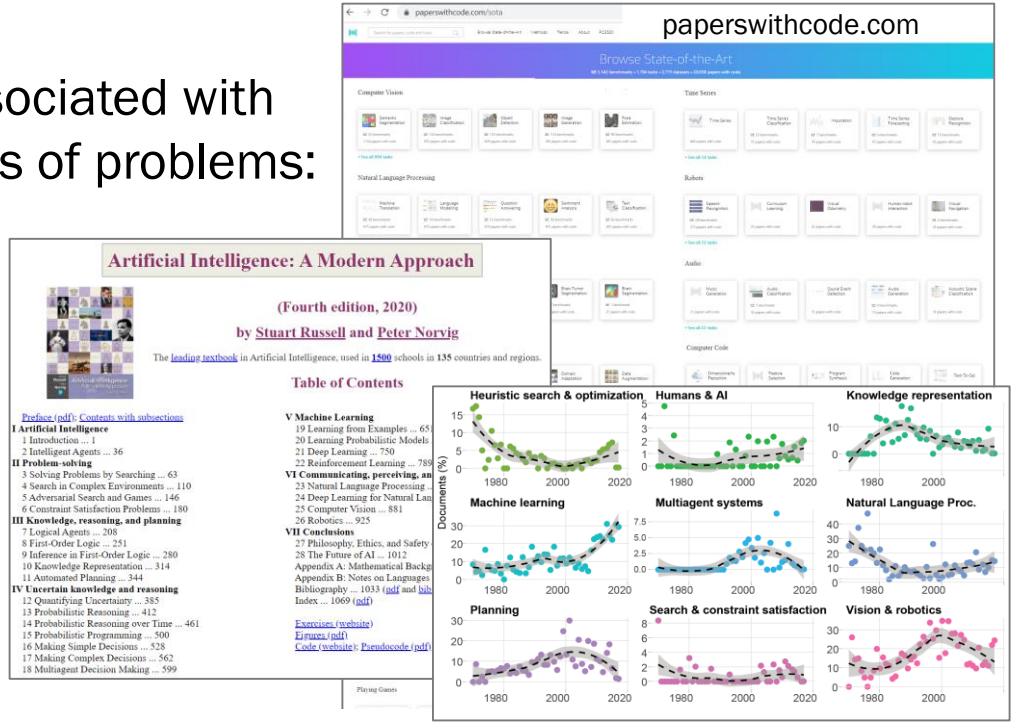




IDENTIFYING CAPABILITIES: TAXONOMIES

AI

- Taxonomies in AI are usually associated with techniques and particular groups of problems:
  - Knowledge Representation
  - Reasoning
  - Planning
  - Learning
  - Perception
  - Navigation
  - Natural Language Processing
  - …



Martínez-Plumed, F., Loe, B. S., Flach, P., O hEigeartaigh, S., Vold, K., & Hernández-Orallo, J. (2018). The facets of artificial intelligence: a framework to track the evolution of AI. In *International Joint Conferences on Artificial Intelligence* (pp. 5180-5187).

IDENTIFYING CAPABILITIES: TAXONOMIES

Pragmatic Integration:

Human tests (From Thurstone to CHC, developmental, cognitive deficit tests, ...)

Animal Cognition (Table of contents of Wasserman and Zentall's book 2006, ...)

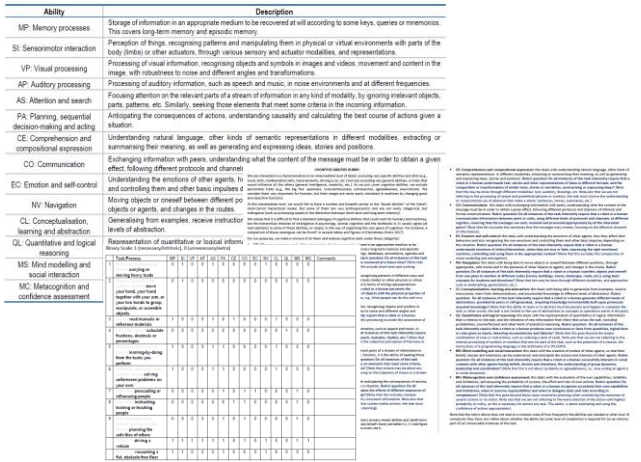
AI (AI textbooks, AI benchmarks, AI Journal, **AGI** categories, ...)

The main criterion for distinguishing two abilities A and B: a system or component (either natural or artificial) could *conceivably* master A but not B.

14 categories

+

a rubric



Ability	Description
MP: Memory processes	Storage of information in an appropriate medium to be recovered at will according to some keys, queries or mnemonics. This covers long-term memory and episodic memory.
SI: Sensorimotor interaction	Perception of things, recognising patterns and manipulating them in physical or virtual environments with parts of the body (limbs) or other actuators, through various sensory and actuator modalities, and representations.
VP: Visual processing	Processing of visual information, recognising objects and symbols in images and videos, movement and content in the image, with robustness to noise and different angles and transformations.
AP: Auditory processing	Processing of auditory information, such as speech and music, in noise environments and at different frequencies.
AS: Attention and search	Focusing attention on the relevant parts of a stream of information in any kind of modality, by ignoring irrelevant objects, parts, patterns, etc. Similarly, seeking those elements that meet some criteria in the incoming information.
PA: Planning, sequential decision-making and acting	Anticipating the consequences of actions, understanding causality and calculating the best course of actions given a situation.
CE: Comprehension and compositional expression	Understanding natural language, other kinds of semantic representations in different modalities, extracting or summarising their meaning, as well as generating and expressing ideas, stories and positions.
CO: Communication	Exchanging information with peers, understanding what the content of the message must be in order to obtain a given effect, following different protocols and channels.
EC: Emotion and self-control	Understanding the emotions of other agents, h… and controlling them and other basic impulses d…
NV: Navigation	Moving objects or oneself between different po… objects or agents, and changes in the routes.
CL: Conceptualisation, learning and abstraction	Generalising from examples, receive instructior… levels of abstraction.
QL: Quantitative and logical reasoning	Representation of quantitative or logical inform…
MS: Mind modelling and social interaction	
MC: Metacognition and confidence assessment	

Hernández-Orallo, J. and K. Vold (2019), AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI, Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). Association for Computing Machinery, New York, NY, USA, 507–513.

Martínez-Plumed, F. et al. (2020), Does AI Qualify for the Job? A Bidirectional Model Mapping Labour and AI Intensities, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA, 94–100.

Ability	Description
MP: Memory processes	Storage of information in an appropriate medium to be recovered at will according to some keys, queries or mnemonics. This covers long-term memory and episodic memory.
SI: Sensorimotor interaction	Perception of things, recognising patterns and manipulating them in physical or virtual environments with parts of the body (limbs) or other actuators, through various sensory and actuator modalities, and representations.
VP: Visual processing	Processing of visual information, recognising objects and symbols in images and videos, movement and content in the image, with robustness to noise and different angles and transformations.
AP: Auditory processing	Processing of auditory information, such as speech and music, in noisy environments and at different frequencies.
AS: Attention and search	Focusing attention on the relevant parts of a stream of information in any kind of modality, by ignoring irrelevant objects, parts, patterns, etc. Similarly, seeking those elements that meet some criteria in the incoming information.
PA: Planning, sequential decision-making and acting	Anticipating the consequences of actions, understanding causality and calculating the best course of actions given a situation.
CE: Comprehension and compositional expression	Understanding natural language, other kinds of semantic representations in different modalities, extracting or summarising their meaning, as well as generating and expressing ideas, stories and positions.
CO: Communication	Exchanging information with peers, understanding what the content of the message must be in order to obtain a given effect, following different protocols and channels of informal and formal communication.
EC: Emotion and self-control	Understanding the emotions of other agents, how they affect their behaviour and also recognising the own emotions and controlling them and other basic impulses depending on the situation.
NV: Navigation	Moving objects or oneself between different positions, through appropriate, safe routes and in the presence of other objects or agents, and changes in the routes.
CL: Conceptualisation, learning and abstraction	Generalising from examples, receive instructions, learn from demonstrations, and accumulate knowledge at different levels of abstraction.
QL: Quantitative and logical reasoning	Representation of quantitative or logical information that is intrinsic to the task, and the inference of new information from them that solves the task, including probabilities, counterfactuals and other kinds of analytical reasoning.
MS: Mind modelling and social interaction	Creation of models of other agents, so that their beliefs, desires and intentions can be understood, and anticipate the actions and interests of other agents.
MC: Metacognition and confidence assessment	Evaluation of the own capabilities, reliability and limitations, self-assessing the probability of success, the effort and risks of own actions.

TESTS: HUMANS

- **Psychometric tests** for general abilities, most notably those related to IQ tests, and other cognitive tests:
  - e.g., WAIS and many others.
- **Developmental tests**: covering a series of stages, sometimes used for various purposes (e.g., detecting mental disabilities):
  - e.g., the Bayle scales, Mullen scales (MSEL), …
- **Tests for general education skills or consolidated knowledge**: exploring "attainment" or "achievement" (often with transversal and basic skills too),
  - e.g., military psychometric tests (ASVAB), college entrance exams (ACT and SAT), vocational educational and training (VET tests), professional (Bennett Mechanical Comprehension Test, BMCT), …

TESTS: AI BY ASKING HUMAN EXPERTS

- **Asking humans:**

Hernández-Orallo, J. "Beyond the Turing Test" Journal of Logic, Language and Information, 2000.

Hernández-Orallo, J. "Twenty Years Beyond the Turing Test: Moving Beyond the Human Judges Too" Minds & Machines, 2020.

  - **Turing Test:** not used in practice, except variants (e.g., CAPTCHAs):
  - **Rubrics:** based on human assessment about AI's capabilities.
    - Using subject matter experts on test questions (e.g., PIAAC).

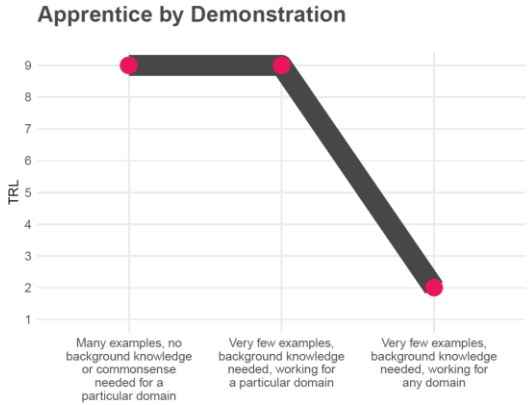
Elliot, S. "Computers and the Future of Skill Demand", OECD 2017

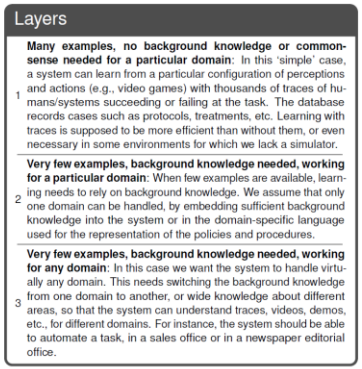
    - Meta-rubrics, can ML automate a task?

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530-1534.

    - TRLs:

More generality → lower TRL





Martínez-Plumed et al., "Futures of Artificial Intelligence through Technology Readiness Levels" under review, 2020

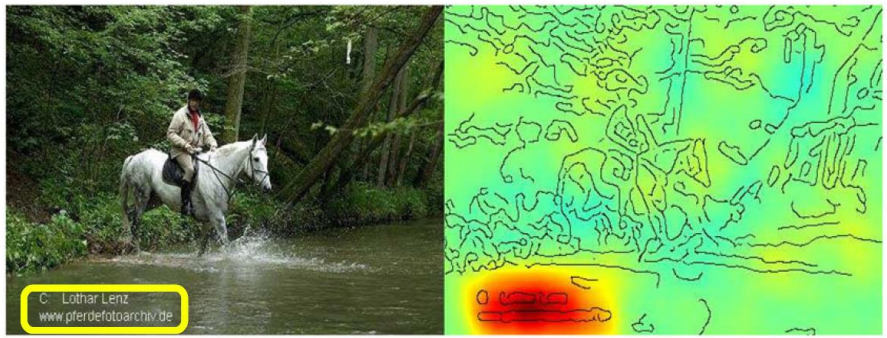
TESTS: AI BY TESTING THE SYSTEM

- Testing the system:
  - **Peer confrontation:** RoboCup, Chess, Go, Poker, etc.,
  - **Benchmarks:** repositories of instances/tasks as challenges for AI.
    - AI reaches superhuman performance but they do not display the capability,
    - Many benchmarks soon replaced.
    - Clever Hans phenomenon:

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

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

Horse-picture from Pascal VOC data set

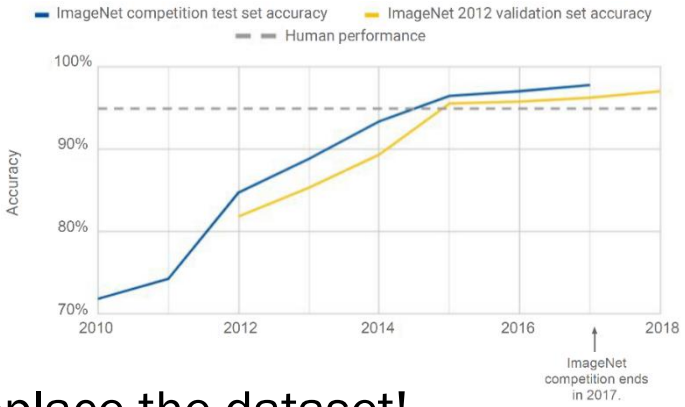


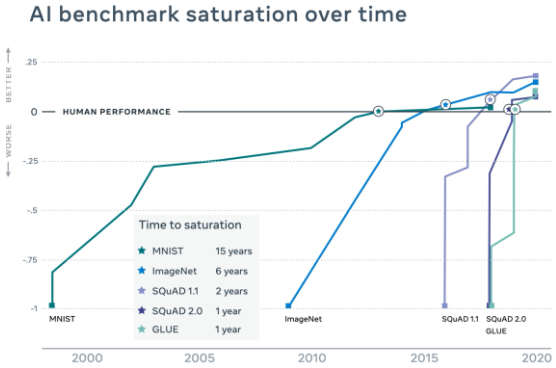
Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1-8.

TESTS: AI NOT ONLY OVERFITTING, ALSO A SCALE PROBLEM

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

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!"

- Replace the dataset!

'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**

TESTS: FROM HUMAN TESTS TO AI?

- Human **tests lack measurement invariance** beyond the human population.
  - These tests are not **proxies** for machines!

Dowe, D. and J. Hernández-Orallo (2012), "IQ tests are not for machines, yet", Intelligence, 40(2).
J Hernández-Orallo, F Martínez-Plumed, U Schmid, M Siebers, DL Dowe (2016) "Computer models solving intelligence test problems: Progress and implications" Artificial Intelligence 230, 74-107

- Humans are agents, while **AI may come as systems and components!**
- Training to the test controlled for humans, but **AI is built on purpose!**
- **Many new capabilities** AI is introducing are **not covered by any human test.**
  - E.g., language identification, generating realistic images, recommendation, …
- Humans and AI differ on the **resources used** (data, compute, sensors) or **external human cognitive labour** (labelling data, human computation).
  - Humans are not allowed to use their extenders but AI can use other AI systems and humans.

Martinez-Plumed, F., Avin, S., Brundage, M., Dafoe, A., hÉigeartaigh, S. Ó., & Hernández-Orallo, J. (2018). Accounting for the neglected dimensions of ai progress. arXiv preprint arXiv:1806.00610.

TESTS: FROM HUMAN/ANIMAL EVALUATION TO AI EVALUATION

- Some hope:
  - Using adaptive testing or adversarial testing,
    - Targeting overfitting (e.g., SWAG in AI2's Mosaic, 

).
  - Item Response Theory and other ideas from psychometrics
    - A populational reference problem! No machine population!
  - Sandbox evaluation: give the elements not the tasks!
    - Let AI researcher build their curricula: then test on unanticipated tasks!
  - Zero-shot, one-shot or few-shot multi-task evaluation (e.g., GPT-3):
    - The same system does different tasks with simple "prompts".

Hernández-Orallo (2020), "Hernández-Orallo, J. "Twenty Years Beyond the Turing Test. Beyond the Human Judges Too" Minds & Machines, 2020.

Martínez-Plumed, F. et al. "Item response theory in AI: Analysing machine learning classifiers at the instance level" *Artificial Intelligence* 271, 18-42, 2019

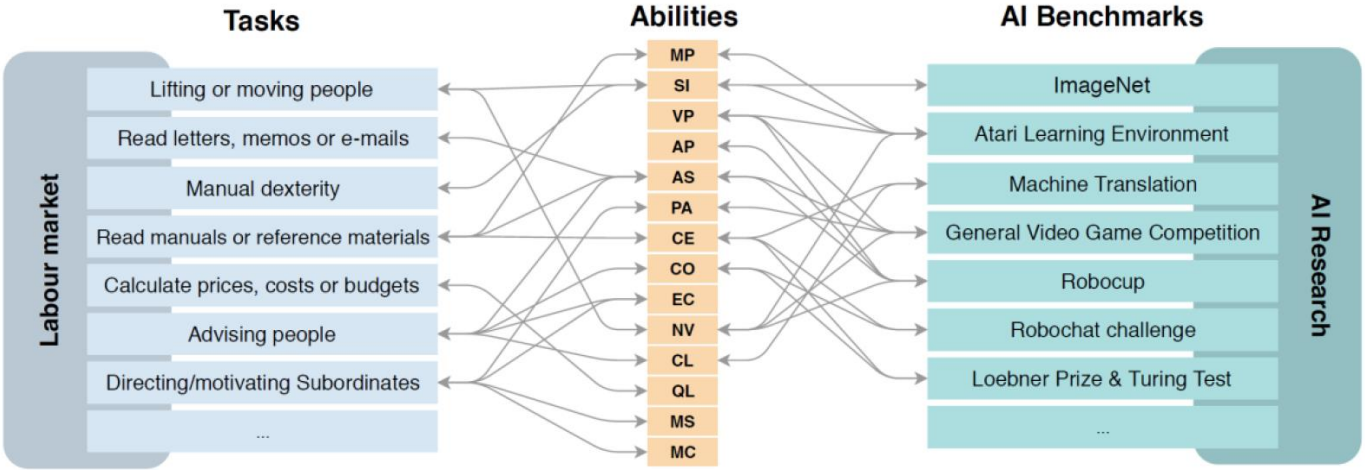
F Martinez-Plumed, J Hernandez-Orallo "Dual indicators to analyse AI benchmarks: Difficulty, discrimination, ability and generality" *IEEE Transactions on Games*, 2020

Crosby, M. et al. (2020), "The animal-ai testbed and competition", PMLR, pp. 164-176.

Many things can be reused from human and animal evaluation, but with stricter Morgan's canons, non-dependence on populations, extra-care in validity, etc.

COMPARISON: THE (INTERMEDIATE) MAPPING APPROACH

- Let's be pragmatic! Can we still compare human tests with AI tests?
  - We can map results through intermediate taxonomies and categories.



Bidirectional and indirect mapping between job market (ISCO-3 specifications) and AI benchmarks

Martínez-Plumed, F. et al. (2020), Does AI Qualify for the Job? A Bidirectional Model Mapping Labour and AI Intensities, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA, 94–100.

http://aicollaboratory.org/

Martinez-Plumed, F. Hernandez-Orallo, J., Gómez, E. "AI Watch: Methodology to Monitor the Evolution of AI Technologies" JRC Working Papers, European Commission, 2020.

Martinez-Plumed, F. Hernandez-Orallo, J., Gómez, E. "Tracking AI: The Capability is (Not) Near", ECAI 2020

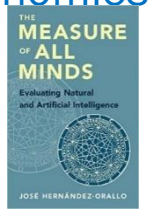
Conclusions

- Skills are changing very rapidly, with extension and collaboration, or displacement, rather than replacement.
- Future AI systems must be less specialised for particular skills and tasks (unless standardised, e.g. driving) featuring abilities and basic skills.
- AI Evaluation has many issues: overfitting, scales, non-autonomy, …
- Tests used in human evaluation do not work for AI, not even as AI becomes more capable, but many concepts can be adapted!
- Common categories and taxonomies are necessary, but we need commensurate scales to appropriately do the mappings.

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/