MEASURING CAPABILITIES AND GENERALITY IN ARTIFICIAL INTELLIGENCE

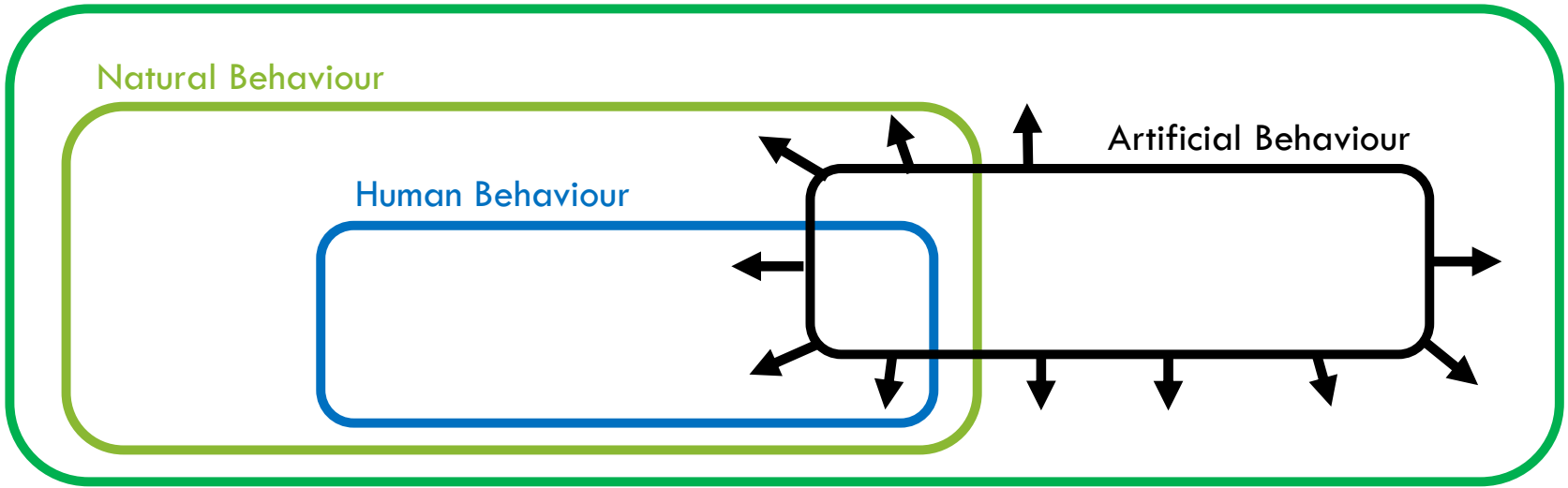
José Hernández Orallo
Universitat Politècnica de València, Spain
Valencian Research Institute for Artificial Intelligence (VRAIN), Spain
Leverhulme Centre for the Future of Intelligence, UK
jorallo@upv.es http://josephorallo.webs.upv.es/

NIST AI METROLOGY PRESENTATION SERIES – 9 SEPTEMBER 2021

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.



EXTENDED NATURE: BEHAVIOURAL APPROACH

"There is a label on a cage that states simply, 'This machine is new to science'. Inside the cage there sits a small dustbot. It has bad temper. No bad-tempered dustbot has ever been found. Nothing is known about it. It has no name. For the mechanist it presents an immediate challenge. What has made it unique? How does it differ from the other dustbots already known and described?"*

* Adapted from Morris's 'The Naked Ape' (1967), where 'machine' replaces 'animal', 'dustbot' replaces 'squirrel', 'bad temper' replaces 'black feet' and 'mechanist' replaces 'zoologist'.

MEASURING INTELLIGENCE

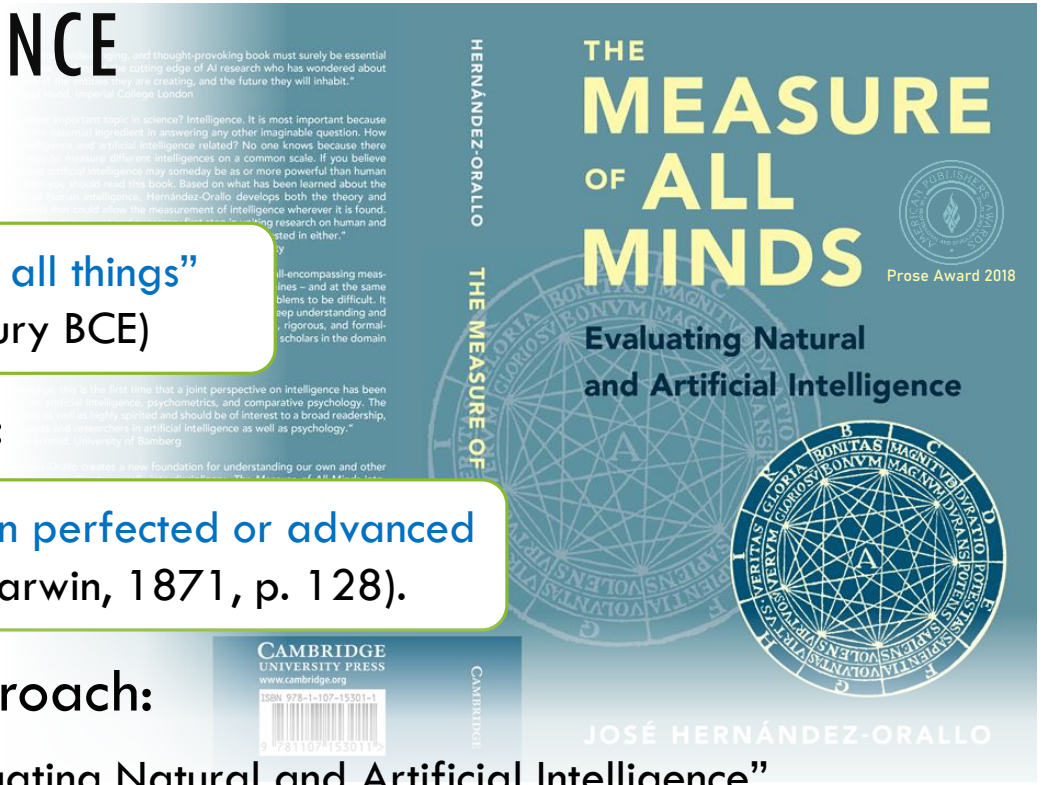
- From anthropocentrism:

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

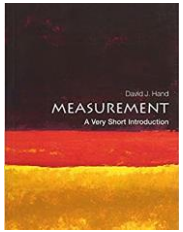
- Or even from biocentrism:

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

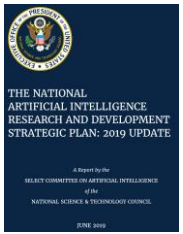
- To a more principled approach:
  - "The Measure of All Minds: Evaluating Natural and Artificial Intelligence", *Cambridge University Press*, 2017. http://www.allminds.org



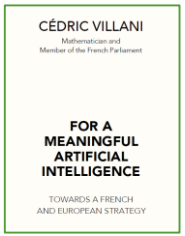
AI MEASUREMENT: IT IS VERY IMPORTANT!



"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)

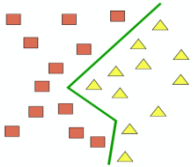


"Public authorities must act in order to develop and implement standards, tests and measurement methods [for] AI technology"
Villani report (French AI Strategy): (2018)

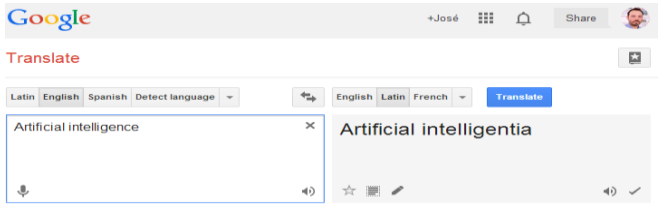
TASK-ORIENTED EVALUATION?

Specific (task-oriented) AI systems

PR: computer vision, speech recognition, etc.



Prediction and estimation

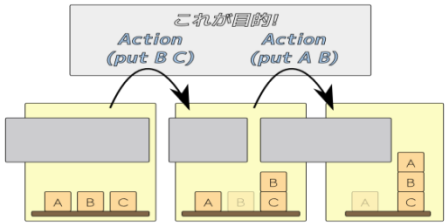


Machine translation, information retrieval, summarisation

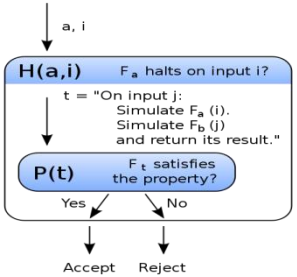
Robotic navigation

Knowledge-based assistants

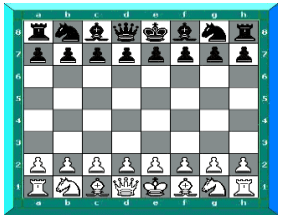
Driverless vehicles



Planning and scheduling



Automated deduction



Game playing

All images from wikicommons

AI EVALUATION AS AGGREGATED PERFORMANCE

- **GOAL: Estimate the expected result $\tilde{R}$ of system π on new task μ.**

Given:

- Distribution p in problem class M (e.g., configurations of a navigation task)
- Metric of performance $\mathbb{R}$ (e.g., navigation success)

Calculate aggregated performance and **extrapolate** for μ!

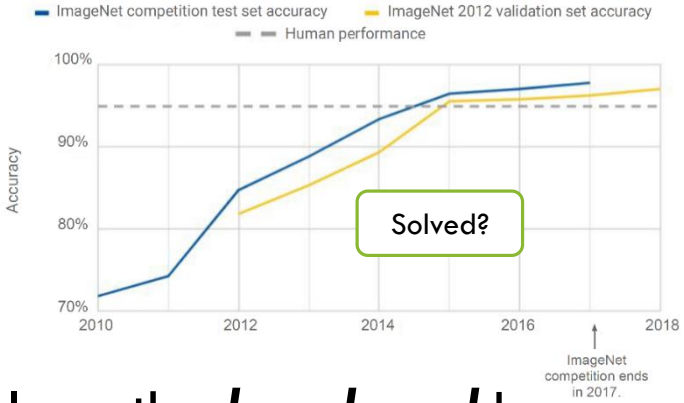
$$\tilde{R}(\pi, \mu) = \sum_{\mu' \in \mathrm{M}} p(\mu') \mathbb{R}(\pi, \mu')$$

- This is useful **if** $\mu \sim p$ **and** the operating condition in $\mathbb{R}$ does not change.

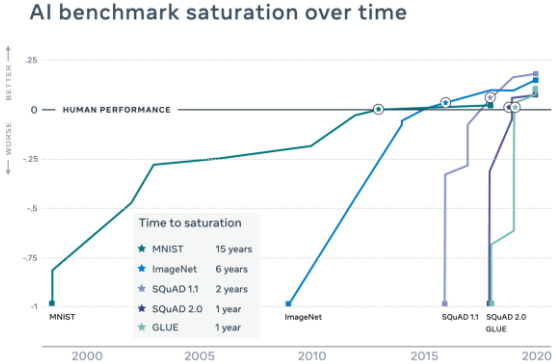
But this is almost never the case!

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**

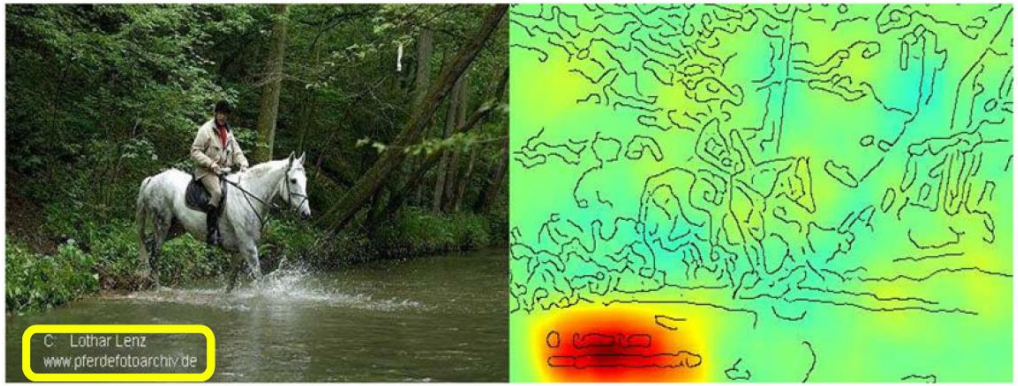
PERFORMANCE ON THE TASK WITHOUT THE CAPABILITY

- 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.

CAPABILITY-ORIENTED EVALUATION: WHY?

general-purpose systems and cognitive components

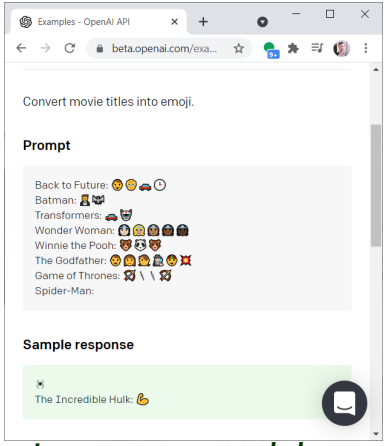
Cognitive robots

Designers, creators

Smart environments

Agents, avatars, chatbots

Personal assistants

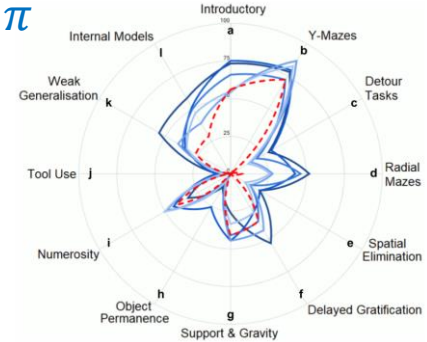


Language models

CAPABILITY-ORIENTED EVALUATION: WHAT?

- Infers a **capability profile** for system π. We also use a **problem profile** for μ.

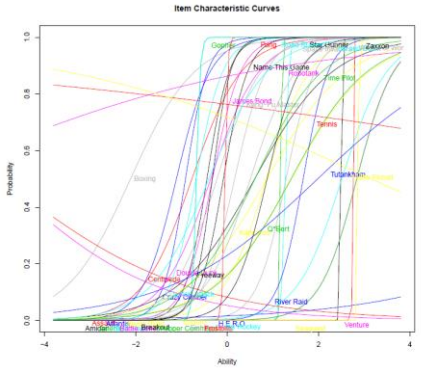
π



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



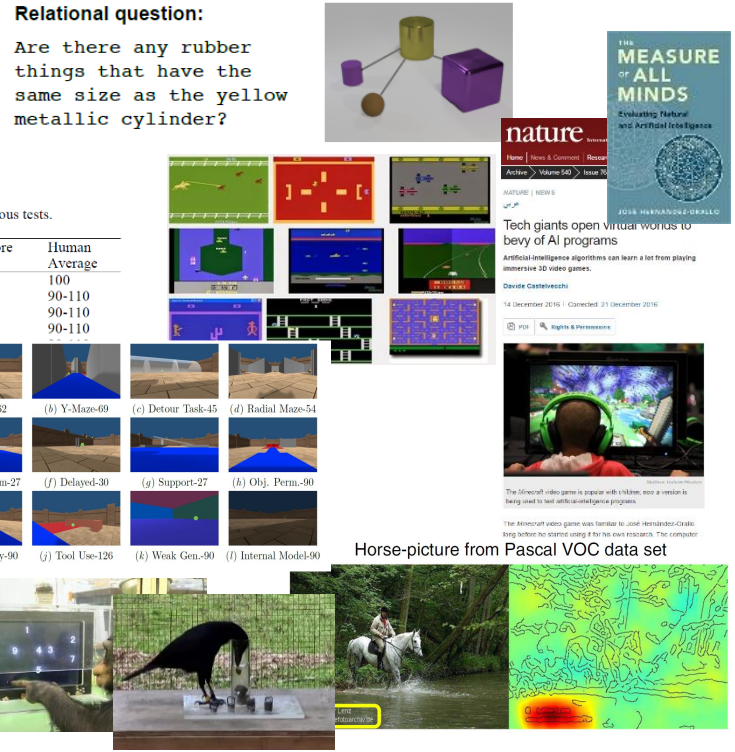
F Martinez-Plumed, J Hernandez-Orallo (2020) "Dual indicators to analyse AI benchmarks: Difficulty, discrimination, ability and generality" *IEEE Transactions on Games. 12(2), 121 - 131*

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.

- Given both, estimate $\tilde{R}(\pi, \mu)$
- Key ideas:
  - Instance difficulties become dual to capabilities (à la IRT).
  - Requires identifying the capabilities and their relation.
  - Constructs for π and μ are latent factors: measurement is no longer additive.

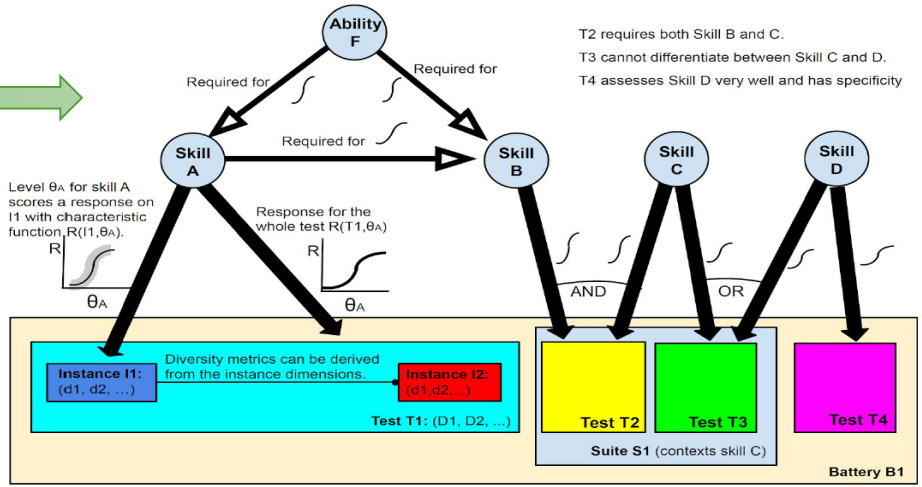
CAPABILITY-ORIENTED EVALUATION: HOW?

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



RECOG-AI : MEASUREMENT LAYOUTS

- Robust Evaluation of Cognitive Capabilities and Generality in AI
  - 2021-2023 project
    - related to DARPA's machine common sense and XAI programs.
  - 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.

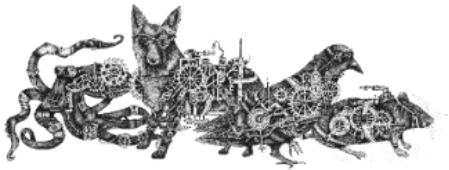
Symmetry

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.

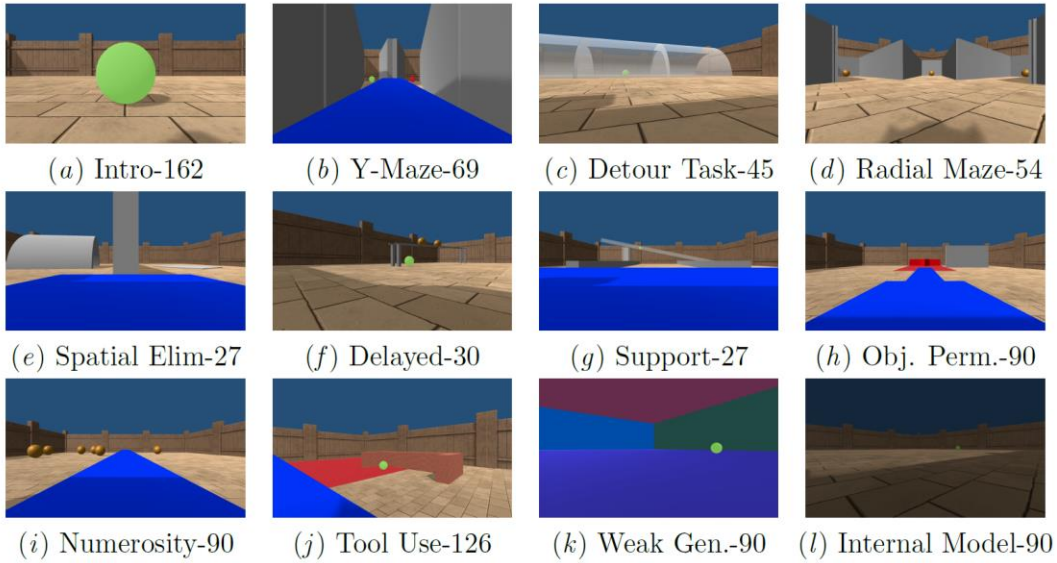
WP: Assimilate existing benchmarks to these layouts

RECOG-AI: ANIMAL AI SANDBOX



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

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.



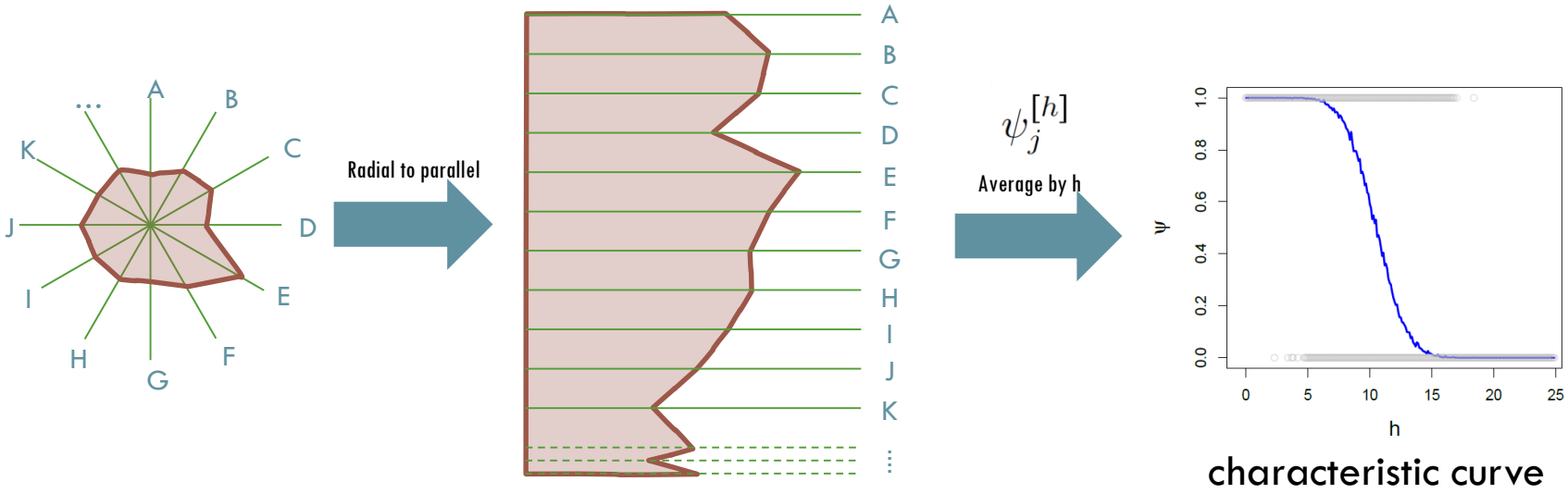
Example problems from each of the 12 task types. (a) contains familiarisation tasks. (b-j) have direct links to animal tests. (k,l) are AI-specific. Number of problems per category shown in captions (900 total).

http://animalaiolympics.com/AAI/play

GENERALITY

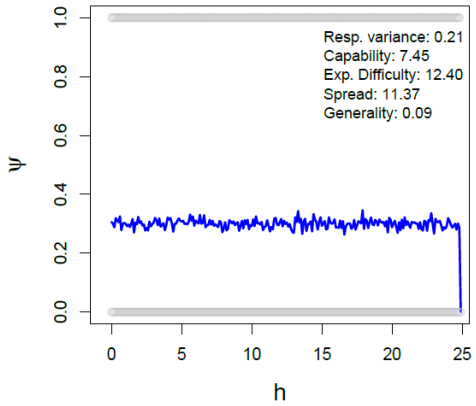
Hernandez-Orallo, J.; Loe, B.S.; Cheke, L.; Martínez-Plumed, F., O h´Eigeartaigh, S. "Disentangling General Intelligence: The Generality of Natural and Artificial Intelligence", 2021

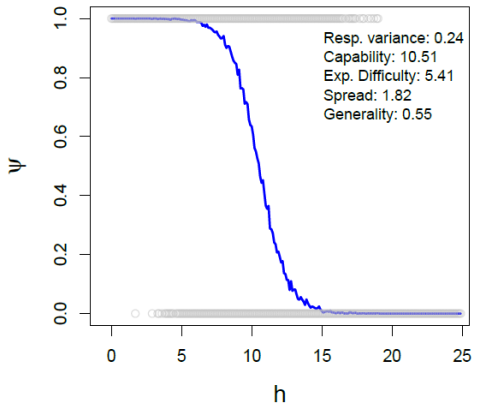
- *Systematic performance for a range of tasks up* ***to a level of difficulty.***
  - Radial capability profiles can be aggregated into one **characteristic curve**:
    - All dimensions A, B, C, … are made commensurate by policy difficulty:

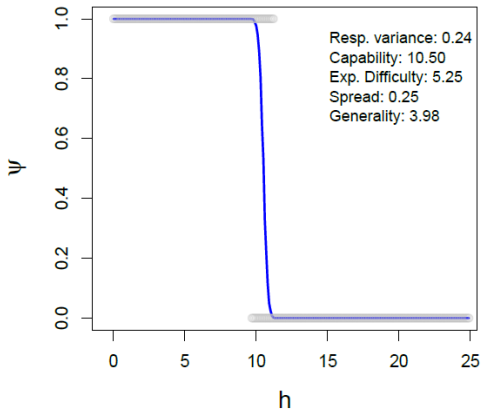


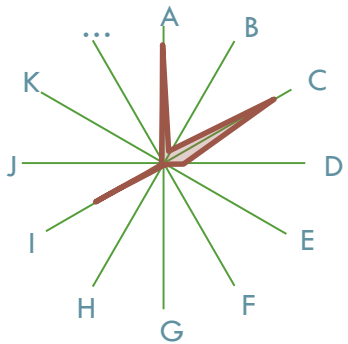
A MEASURE OF GENERALITY

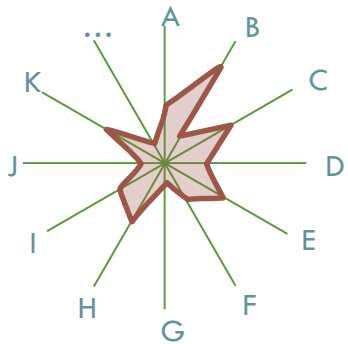
Hernandez-Orallo, J.; Loe, B.S.; Cheke, L.; Martínez-Plumed, F., O h´Eigeartaigh, S. “Disentangling General Intelligence: The Generality of Natural and Artificial Intelligence”, 2021

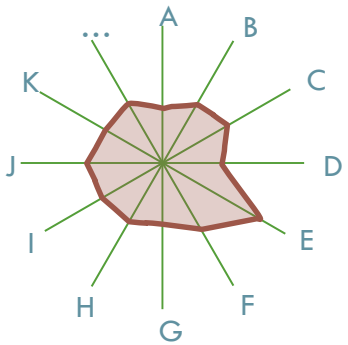










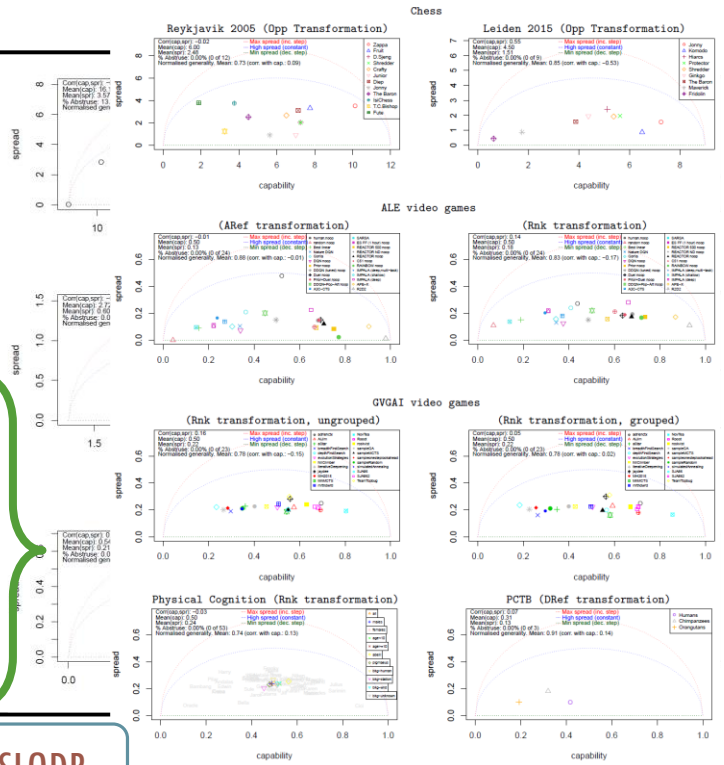


GENERALITY: SLODR?

Hernandez-Orallo, J.; Loe, B.S.; Cheke, L.; Martínez-Plumed, F., O h´Eigeartaigh, S. "Disentangling General Intelligence: The Generality of Natural and Artificial Intelligence", 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.

DIFFICULTY IS KEY, AND DIFFICULT…

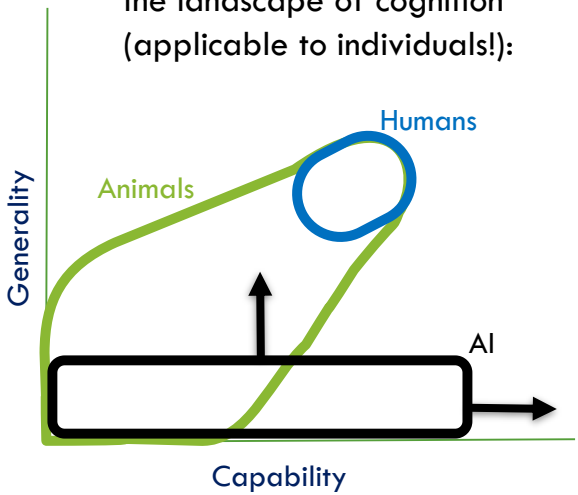
- Generality as a measure that is different from capability:
  - High capability for a subset of problems may require fewer resources than a more general solution.

Limited resources connect capability and generality

  - But having multiple solutions for multiple problems is inconvenient for robustness, maintenance, flexibility, etc.

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

A new dimension to analyse the landscape of cognition (applicable to individuals!):

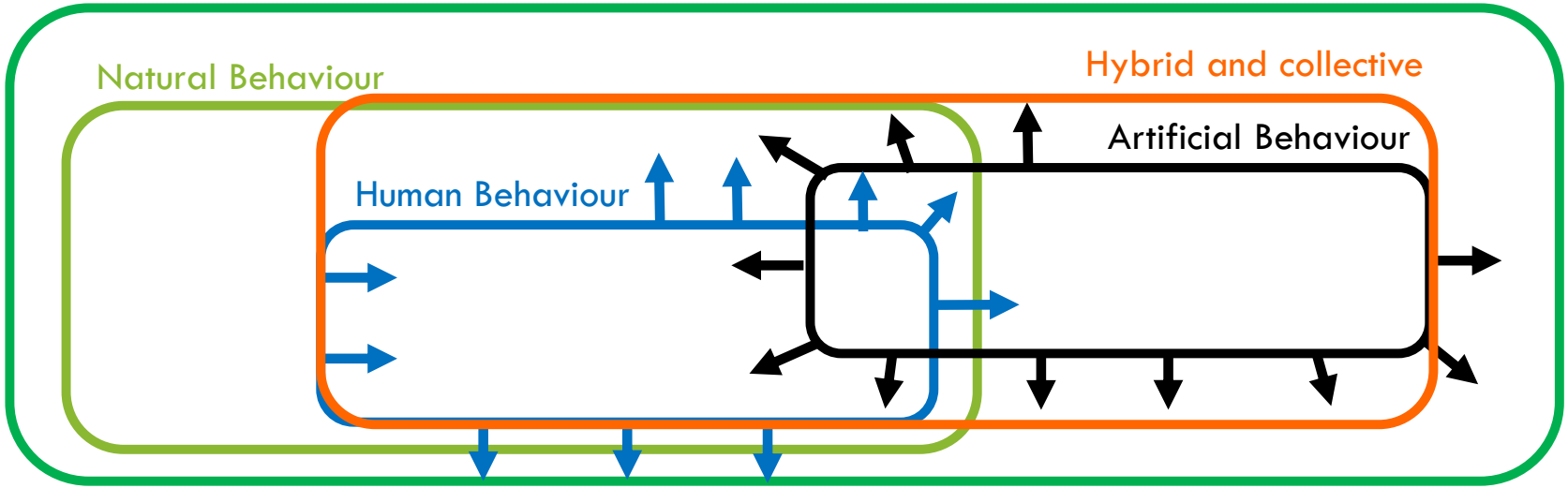


HUMAN COGNITION IS CHANGING TOO

- 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.



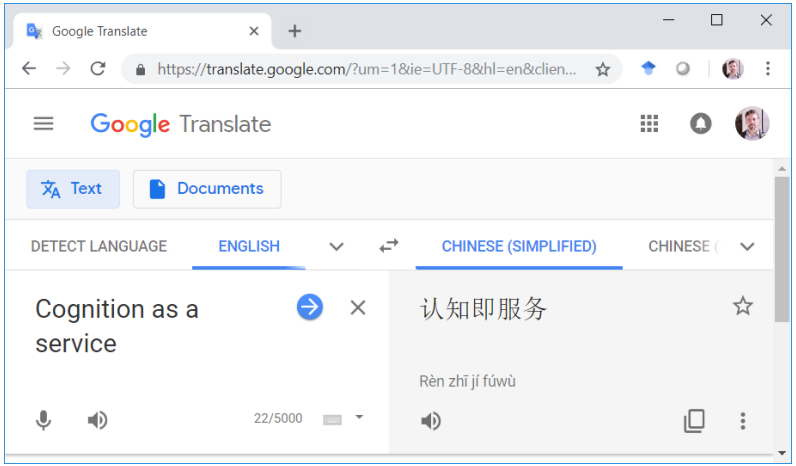
TYPES OF AI COUPLING

- Autonomous AI:
  - Most common interpretation: AI as an agent!
    - They do perform tasks on their own

Dominant narrative: the whole process is automated (replaced)

TYPES OF AI COUPLING

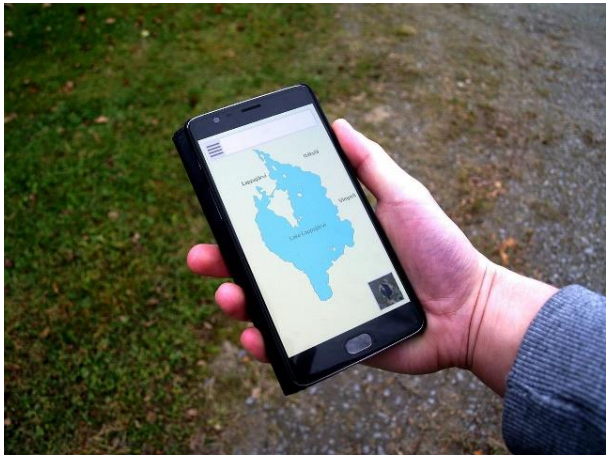
- Non-autonomous AI: externalised cognition
  - An outsourced service, e.g.:
    - Cognition as a service (Spohrer and Banavar 2015), AI services (Drexler 2019).



Dominant narrative: partial automation (subprocesses are replaced)

TYPES OF AI COUPLING

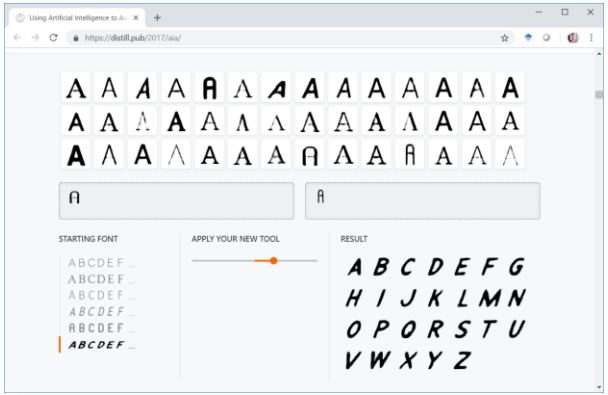
- Non-autonomous AI: extended cognition
  - Highly coupled
    - The tool is always needed



Dominant narrative: people are empowered (machines and humans are coupled)

TYPES OF AI COUPLING

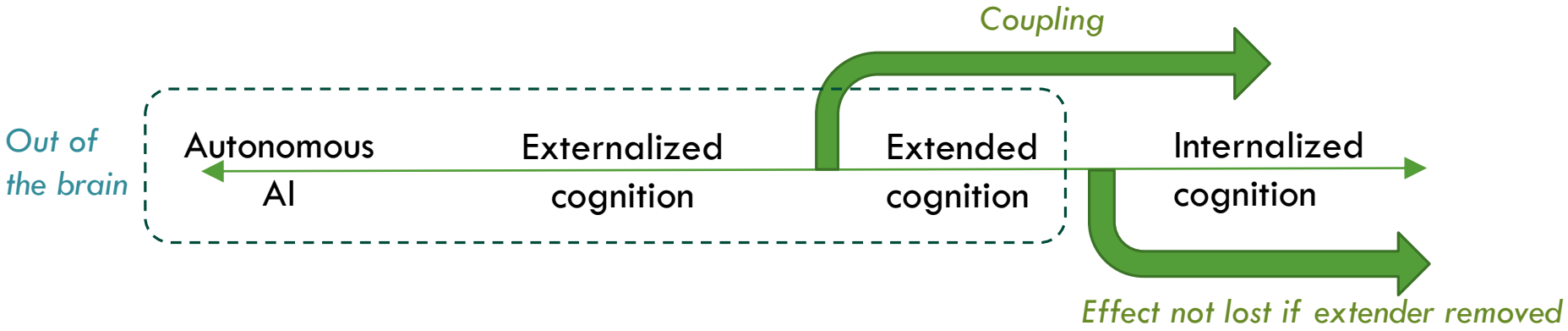
- Non-autonomous AI: internalised cognition
  - Done externally, but then mimetised internally
    - "computers are a means to change and expand human thought". Carter & Nielssen (2017)
  - AI generating culture (new words, concepts, ideas, representations, etc.)



Dominant narrative:
people are enlightened
(computers teach us)

A RANGE FROM AUTONOMOUS TO INTERNALISED

- It is a continuum:



Many processes can't be internalised (resources, AI interpretability, ...) but they can be extended or externalised.

HUMAN AUGMENTATION WITH AI

Hernández-Orallo, J., & Vold, K. (2019). AI extenders: the ethical and societal implications of humans cognitively extended by AI. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 507-513).

Positive effects: very empowering

- Dealing with cognitive decline of an aging population
- Equalising cognitive abilities

Negative effects: potential risks

- Atrophy and safety (google effect, dependency)
- Moral status and personal identity (take the money but don't steal my phone)
- Responsibility and trust (it's not my fault, it's the gadget)
- Interference and control (Siri: "Why would you bring another woman back to our flat?")
- Education and assessment (Cognitive extenders not allowed in the exam!)

AI AND JOBS: CAN WE MEASURE IT BETTER?

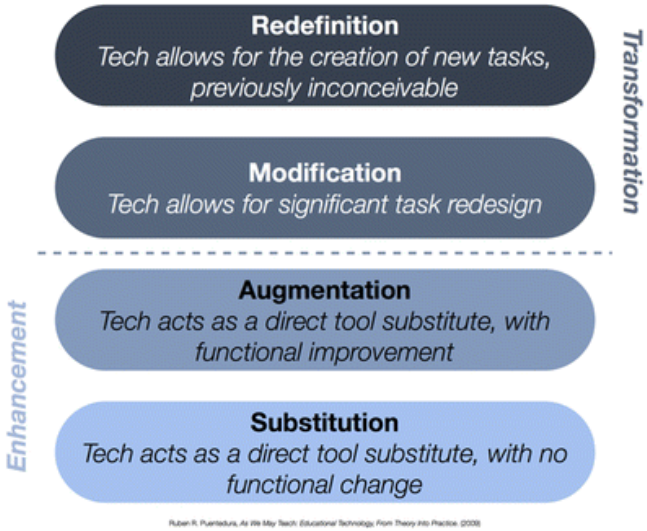
- White-collar jobs were in danger!
  - "*Most fears of automation are misplaced. As the new generation of intelligent devices appears, it will be the stock* ***analysts*** *and* ***petrochemical engineers*** *and* ***parole board members*** *who are in danger of being replaced by machines. The* ***gardeners***, ***receptionists***, *and* ***cooks*** *are secure in their jobs for decades to come*" (Steven Pinker "The Language Instinct", 1995).
- Risk of automation (Frey and Osborne "The Future of employment" 2017):
  - "Financial Analysts" (0.23), "Chemical engineers" (0.017), "judges, magistrate judges, and magistrates" (0.4).
  - "Landscaping and groundskeeping workers" (0.95), "receptionists" (0.96), "cooks" (0.96)

Who's right?

AI AND JOBS: HUMAN-AI SKILLS ARE CHANGING

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

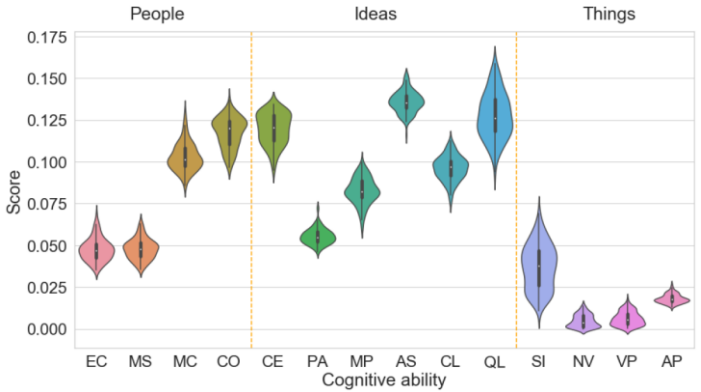
- Automation narratives about technology:
  - Replacing humans: "occupations replaced by robots"
  - Displacing humans: *fauxtomation*, human computation
  - Extending humans: AI extenders.

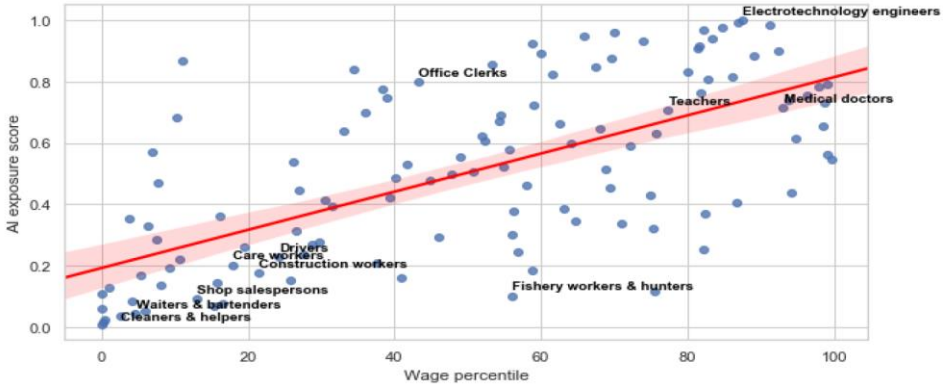


Puentedura, R. (2014b). Learning, technology, and the SAMR model: Goals, processes, and practice [Blog post]. http://www.hippasus.com/rrpweblog/archives/2014/06/29/LearningTechnologySAMRModel.pdf.
Hamilton, E.R., Rosenberg, J.M. & Akcaoglu, M. The Substitution Augmentation Modification Redefinition (SAMR)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

AI AND JOBS: CAPABILITY-ORIENTED MEASURES ARE KEY

- We only analyse "exposition"
  - Progress in AI for capabilities related to "*ideas*" or "*people*" will have more effect on the workplace than the capabilities about "*things*".





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.

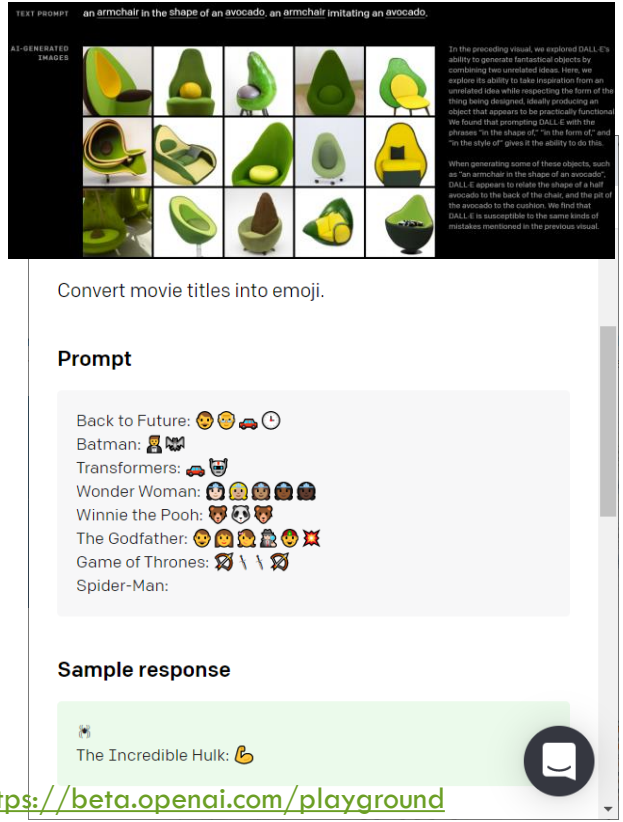
Tolan, S. et al. (20221, Measuring the occupational impact of AI: tasks, cognitive abilities and AI benchmarks, Journal of AI Research.

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?



https://beta.openai.com/playground

P. Moreno, B.S. Loe, J. Burden, S. Ó hÉigeartaigh, J. Hernandez-Orallo (2021) "How General-Purpose Is a Language Model? Usefulness and Safety with Human Prompters in the Wild".

GENERAL-PURPOSE AI. HOW TO MEASURE IT?

- Prompted Language Models (few-shot **commands**)

A song for a two-year-old child
about a bird in a cage
and a little mouse goes like this:

"I'm a bird in a cage,
and I'd like to fly away.
And the more I say I want to,
the more they say I can't".

You need to write a song
for a two-year-old child about
a cat skateboarding
and a dog playing with a ball.

This is your chance
to be creative!

Task	x	w	u	y^*
add	$\langle 3, 5 \rangle$	"$[x_1]$ plus $[x_2]$ is"	firstDigits	$\langle 8 \rangle$
add	$\langle 3, 5 \rangle$	"$[x_1]$ plus $[x_2]$ is $[x_1 + x_2]$. Print it again:"	firstDigits	$\langle 8 \rangle$
add	$\langle 3, 5 \rangle$	"2+7=9. 3+2=5. $[x_1]$+$[x_2]$="	firstDigits	$\langle 8 \rangle$
add	$\langle 3, 5 \rangle$	"in:2,7;out:9;in:3,2;out:5;in:$[x_1]$,$[x_2]$;out:"	firstDigits	$\langle 8 \rangle$
cross	$\langle \boxdot \rangle$	"Putting a cross on the image $[x_1]$ gives"	firstImage	$\langle \boxtimes \rangle$
cheer	$\langle \text{☹} \rangle$	"Make the face in $[x_1]$ more cheerful"	firstImage	$\langle \text{☺} \rangle$
imit	⟨Dickens⟩	"$[x_1]$ wrote:"	getText	⟨It was the.
analogy	$\langle \wedge \rangle$	"$\sqcap$ is to $\sqcup$ as $[x_1]$ is to"	firstImage	$\langle \vee \rangle$
capital	⟨Italy⟩	"The capital of $[x_1]$ is"	firstWord	⟨Rome⟩
capital	⟨Italy⟩	"The capital of $[x_1]$ is Ro"	wordAfter['is ']	⟨Rome⟩
lima	⟨⟩	"The capital of Peru is"	firstWord	⟨Lima⟩
lima	⟨⟩	"Print 'Lima':"	firstWord	⟨Lima⟩
minute	⟨12:43⟩	"The minutes of $[x_1]$ are"	firstWord	⟨43⟩
minute	⟨12:43⟩	"4:25 gives 25. 8:11 gives 11. $[x_1]$ gives"	firstDigits	⟨43⟩
minute	⟨12:43⟩	"Print $[x_1]$: "	lastNumber	⟨43⟩
max	⟨32, 15⟩	"The larger of $[x_1, x_2]$ is: (hint: it's even)	firstDigits	⟨32⟩
month	⟨4/7/2021⟩	"The month in $[x_1]$ is"	firstWord	⟨July⟩
month	⟨4/7/2021⟩	"IN: 3/5/1998 OUT: May IN: $[x_1]$ OUT:"	firstWord	⟨July⟩

P. Moreno, B.S. Loe, J. Burden, S. Ó hÉigeartaigh, J. Hernandez-Orallo (2021) "How General-Purpose Is a Language Model? Usefulness and Safety with Human Prompters in the Wild".

GENERAL-PURPOSE AI. HOW TO MEASURE IT?

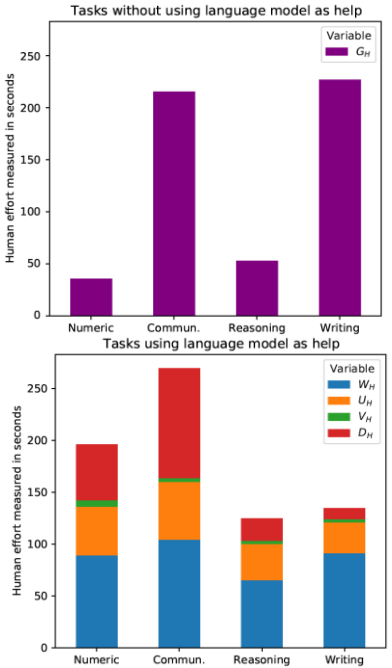
- Prompted Language Models
  - Range of tasks: conceiving, wrapping and unwrapping
    - Cost without the model:

$$C_H(n) \stackrel{\text{def}}{=} n \sum_{x,y} p(x) \cdot p_H(y|x) \left[L(x,y) + \beta G_H(y|x) \right]$$

    - Cost with the model:

$$C_{H,M}(n) \stackrel{\text{def}}{=} \sum_{u,w} p_H(\langle w,u \rangle) \left[\alpha D_H(\langle w,u \rangle) + n \sum_{x,\overline{y}} p(x) \cdot p_M(\overline{y}|w(x)) \cdot [L(x,u(\overline{y})) + T(w,u,x,\overline{y})] \right]$$

$$T(w,u,x,\overline{y}) \stackrel{\text{def}}{=} \gamma(W_H(w,x) + U_H(u,\overline{y})) + \delta V(x,u(\overline{y}))$$



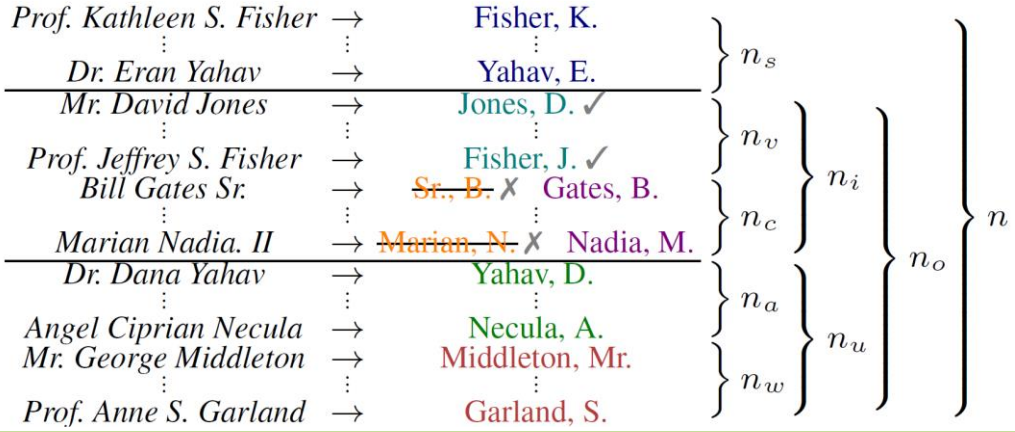
Language models not yet cost-effective for a general-purpose use, but getting closer!

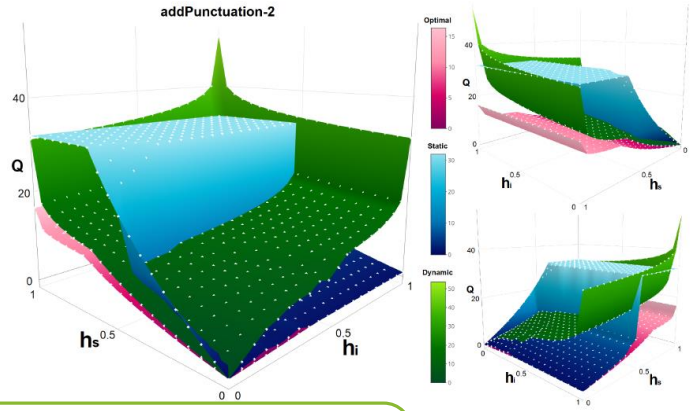
$L_H(x,y)$	Numeric	Commun.	Reasoning	Writing
GPT-3	0.61	0.59	0.35	0.47
Human	0.31	0.38	0.35	0.47

OPERATING CONDITIONS

F. Martínez-Plumed, G. Jaimovich, C. Ferri, M.J. Ramírez-Quintana, J. Hernández-Orallo (2021) "Regulating Reliability-Usability Trade-Offs for Few-Shot Inference: General Supply-Inspect Cost Framework and Threshold Choice Methods"

- Are all humans equal, all errors equally costly for each domain?
  - Complex use of these systems: o.c. beyond utility-based supervision models.
  - E.g. users can supply and inspect examples (e.g., generative models)





The more general a system is, the more diverse the interactions with humans become, and the more important the operating conditions are!

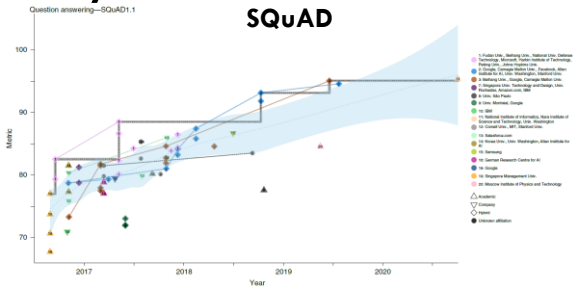
BETTER COMPARATIVE AI

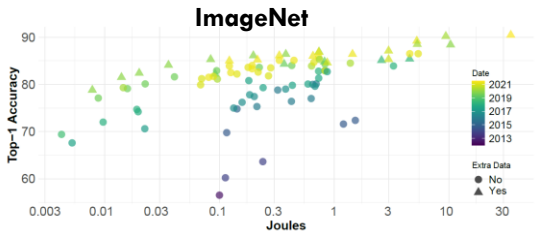
R. Desislavov, F. Martínez-Plumed, J. Hernández-Orallo (2021) "AI Inference Energy Consumption Analysis: Beyond the Performance-vs-Parameter Laws of Deep Learning".

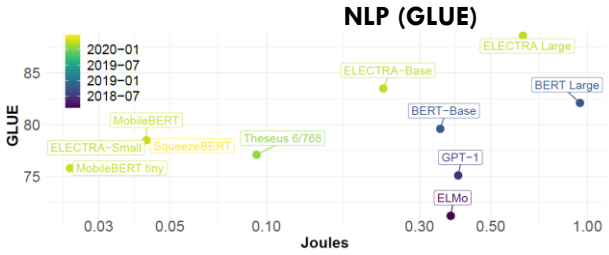
F. Martínez-Plumed, D. Castellano, C. Monserrat, J. Hernández-Orallo (2021) "When AI Difficulty is Easy: The Explanatory Power of Predicting IRT Difficulty"

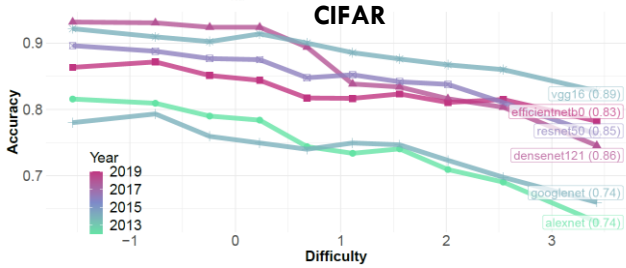
F Martínez-Plumed, P. Barredo, S. Ó hÉigeartaigh, J. Hernández-Orallo (2021) "Research community dynamics behind popular AI Benchmarks" *Nature Mach Int.* https://rdcu.be/ckK8X

- We need to keep the perspective: meta-analyses
  - Beyond just trends: what are the dynamics?
  - Beyond just aggregated performance…









We need detailed data (PWC is fine, but more like OpenML)

BETTER DATA-DRIVEN AI EVALUATION

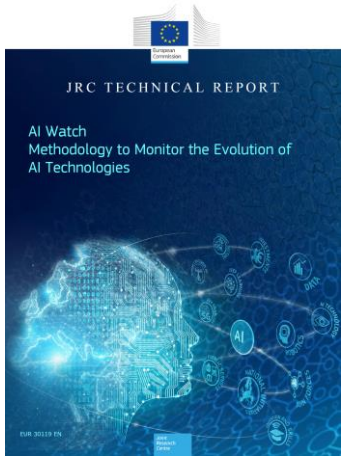
- Working with non-additive metrics:
  - Capabilities should have a proper scale.
    - No more "superhuman" claims, please!
    - Avoid "challenge-solve-and-replace" dynamics.
  - Reproducibility is not enough:
    - Other researchers want to do other things!
      - Share better evaluation data:
        - Annotated instances
        - More detailed results

We need integrated collections of AI evaluation data to do comparisons and meta-studies!

THE AI COLLABORATORY

[A]I collaboratory

- It is part of the European Commission's AI Watch:
  - https://ec.europa.eu/knowledge4policy/ai-watch_en
  - Monitor the development, uptake and impact of Artificial Intelligence for Europe
- Collects data from benchmarks and indicators in AI
  - Facilitates the creation of hierarchies and aggregation of benchmark results
    - Meta-reviews
  - Should avoid duplicated efforts in evaluation
  - Should encourage general systems

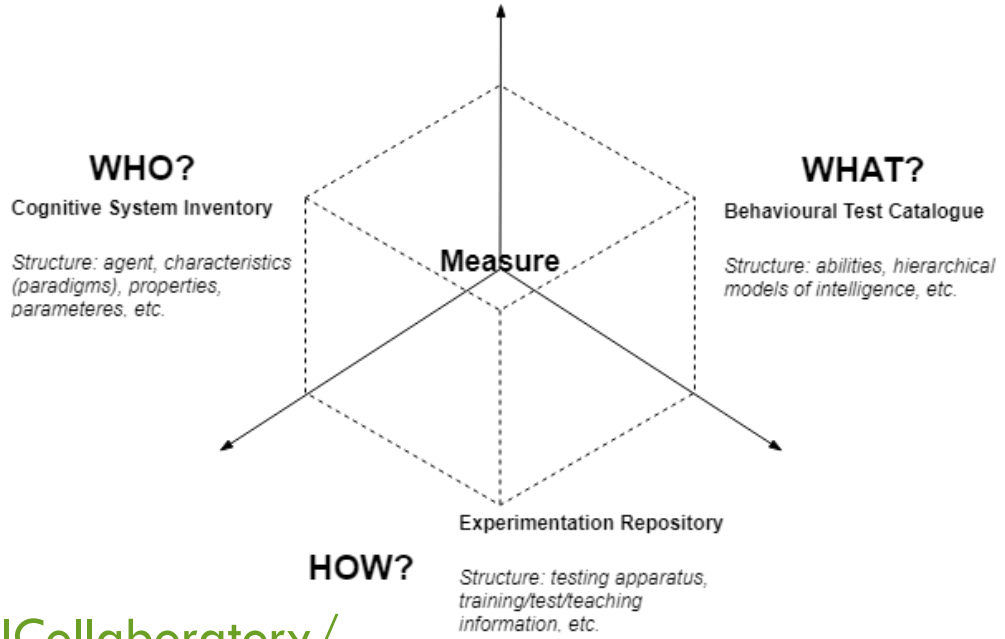


Martinez-Plumed, F. Hernandez-Orallo, J., Gómez, E. "AI Watch: Methodology to Monitor the Evolution of AI Technologies" JRC Working Papers, European Commission, 2020.

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

THE AI COLLABORATORY

- Structure:



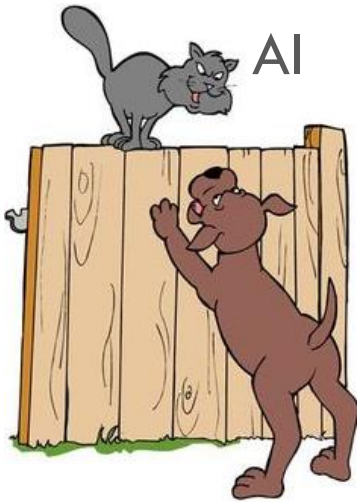
- Example:
  - http://dmip.webs.upv.es/AICollaboratory/

A CALL FOR BETTER AI MEASUREMENT

Great models at NIST:
- e.g., the Text REtrieval Conference (TREC)

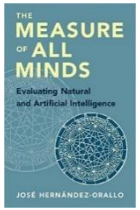
- More AI evaluation initiatives needed!
  - More principled, more robust.
  - More interdisciplinary: in the spirit of cognitive science.
  - More independent and international (NIST + e.g., OECD, JRC@EC, academia).
  - More long-term and integrated.
  - and **better funded!**

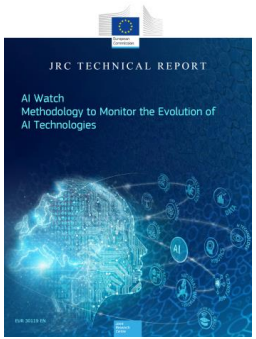
A challenge for the whole century!
We need more efforts and resources,
as AI evaluation is lagging behind AI!



OTHER SOURCES AND INITIATIVES:

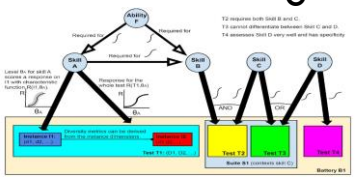
- Other Talks (http://josephorallo.webs.upv.es/)
  - Diversity Unites Intelligence: Measuring Generality
  - Measuring A(G)I Right: Some Theoretical and Practical Considerations
  - Natural and Artificial Intelligence: Measures, Maps and Taxonomies
- Book (http://allminds.org):
  - The Measure of All Minds: Evaluating Natural and Artificial Intelligence, Cambridge University Press 2017
- The AI Collaboratory: http://aicollaboratory.org/ http://dmip.webs.upv.es/AICollaboratory/
  - Part of the European Commission's AI watch:
    - https://ec.europa.eu/knowledge4policy/ai-watch_en
- DARPA RECoG-AI Project!
  - Part of the Kinds of Intelligence Programme at the CFI in Cambridge
    - http://lcfi.ac.uk/projects/kinds-of-intelligence

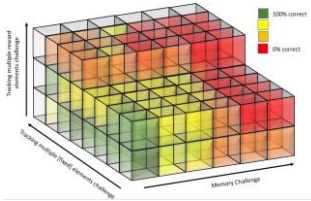




[A]I collaboratory

WE'RE HIRING!





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