Panel 6: Metrics and Measurement Methods:
What and How to Test

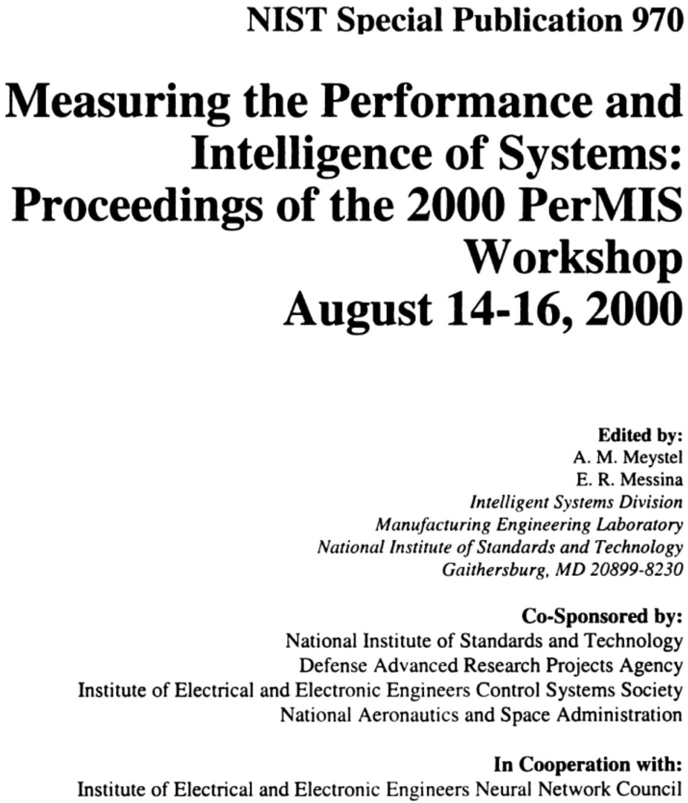
José Hernández-Orallo (jorallo@upv.es)
Valencian Research Institute for Artificial Intelligence (vrAIn) (vrain.upv.es)
Universitat Politècnica de València, València (www.upv.es)
Leverhulme Centre for the Future of Intelligence, Cambridge (lcfi.ac.uk)

VRAIN

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

CFI LEVERHULME CENTRE FOR THE
FUTURE OF INTELLIGENCE

NIST Workshop on AI Measurement and Evaluation, June 15-17, 2021



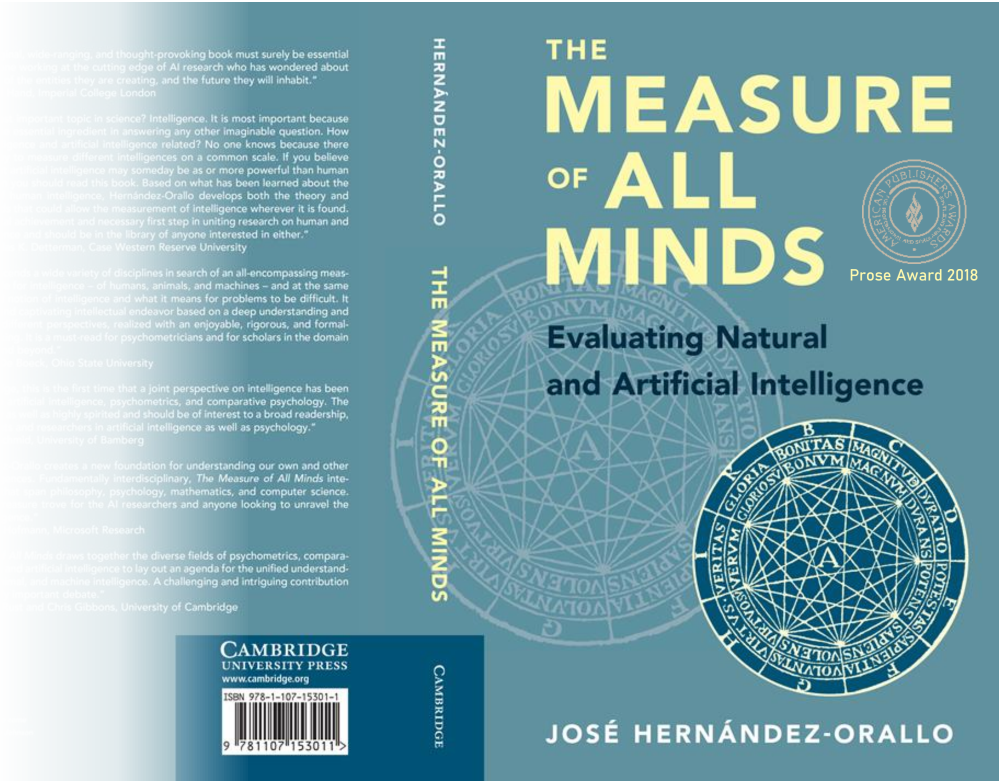
NIST Special Publication 970

Measuring the Performance and Intelligence of Systems: Proceedings of the 2000 PerMIS Workshop August 14-16, 2000

Edited by:
A. M. Meystel
E. R. Messina
Intelligent Systems Division
Manufacturing Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899-8230

Co-Sponsored by:
National Institute of Standards and Technology
Defense Advanced Research Projects Agency
Institute of Electrical and Electronic Engineers Control Systems Society
National Aeronautics and Space Administration

In Cooperation with:
Institute of Electrical and Electronic Engineers Neural Network Council



wide-ranging, and thought-provoking book must surely be essential
working at the cutting edge of AI research who has wondered about
the entities they are creating, and the future they will inhabit."
Hand, Imperial College London

important topic in science? Intelligence. It is most important because
essential ingredient in answering any other imaginable question. How
gence and artificial intelligence related? No one knows because there
to measure different intelligences on a common scale. If you believe
artificial intelligence may someday be as or more powerful than human
you should read this book. Based on what has been learned about the
human intelligence, Hernández-Orallo develops both the theory and
that could allow the measurement of intelligence wherever it is found.
achievement and necessary first step in uniting research on human and
ce and should be in the library of anyone interested in either."
K. Detterman, Case Western Reserve University

a wide variety of disciplines in search of an all-encompassing meas-
for intelligence – of humans, animals, and machines – and at the same
notion of intelligence and what it means for problems to be difficult. It
captivating intellectual endeavor based on a deep understanding and
different perspectives, realized with an enjoyable, rigorous, and formal-
It is a must-read for psychometricians and for scholars in the domain
beyond."
Boeck, Ohio State University

this is the first time that a joint perspective on intelligence has been
artificial intelligence, psychometrics, and comparative psychology. The
well as highly spirited and should be of interest to a broad readership,
and researchers in artificial intelligence as well as psychology."
Schmid, University of Bamberg

Orallo creates a new foundation for understanding our own and other
Fundamentally interdisciplinary, *The Measure of All Minds* inte-
span philosophy, psychology, mathematics, and computer science.
treasure trove for the AI researchers and anyone looking to unravel the
ence."
Hofmann, Microsoft Research

All Minds draws together the diverse fields of psychometrics, compara-
artificial intelligence to lay out an agenda for the unified understand-
and machine intelligence. A challenging and intriguing contribution
important debate."
and Chris Gibbons, University of Cambridge

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-1-107-15301-1

HERNÁNDEZ-ORALLO THE MEASURE OF ALL MINDS CAMBRIDGE

THE MEASURE OF ALL MINDS

Evaluating Natural and Artificial Intelligence

JOSÉ HERNÁNDEZ-ORALLO

Prose Award 2018

AI Evaluation as Aggregated Performance

- GOAL: Estimate the expected result $\tilde{R}$ of system π and a 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!

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

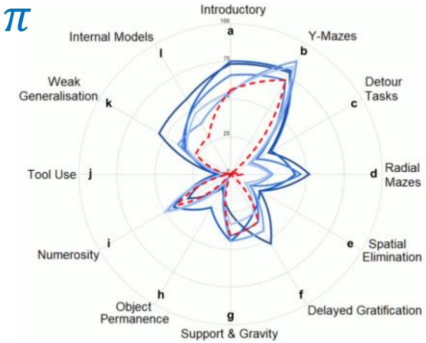
- This is useful **if** $\mu \sim p$ **and** the operating condition in $\mathbb{R}$ does not change.

But this is almost never the case!

From Task-oriented to Capability-oriented Evaluation

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

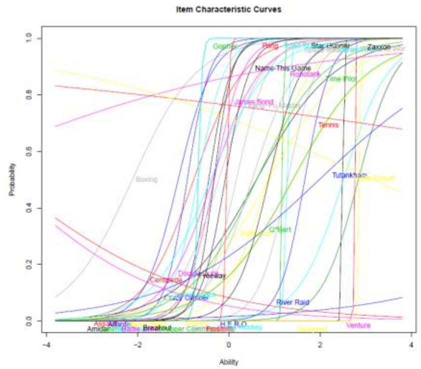
π



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



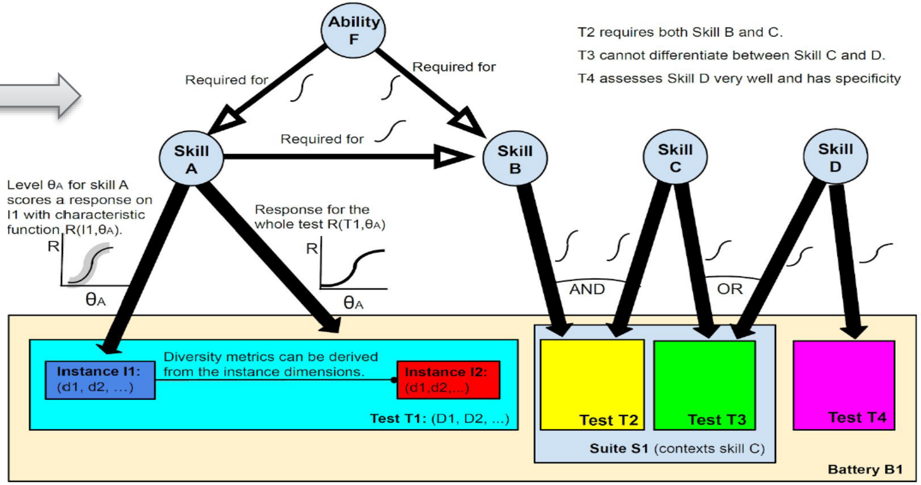
F Martinez-Plumed, J Hernandez-Orallo (2020) "Dual indicators to analyse AI benchmarks: Difficulty, discrimination, ability and generality" *IEEE Transactions on Games. 12(2), 121 - 131*

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 $\tilde{R}(\boldsymbol{\pi}, \boldsymbol{\mu})$
- 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.

RECOG-AI : MEASUREMENT LAYOUTS

- Robust Evaluation of Cognitive Capabilities and Generality in AI
  - 2021-2023 (planning to work with DARPA)
    - related to the machine common sense program, director: Matt Turek.
  - 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



T2 requires both Skill B and C.

T3 cannot differentiate between Skill C and D.

T4 assesses Skill D very well and has specificity

Level θA for skill A scores a response on I1 with characteristic function R(I1,θA).

Response for the whole test R(T1,θA)

Diversity metrics can be derived from the instance dimensions.

RECOG-AI : SPACES AND FEATURES

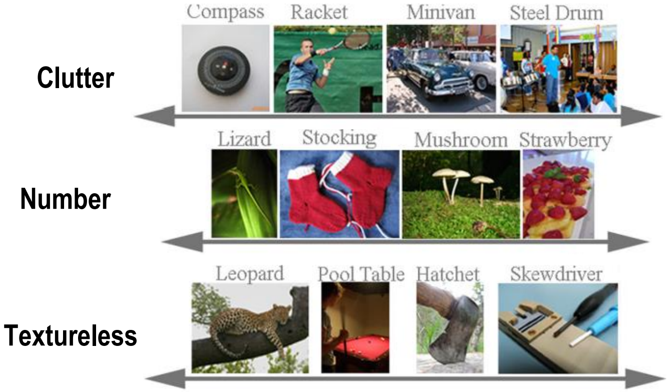
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211-252.

- 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.

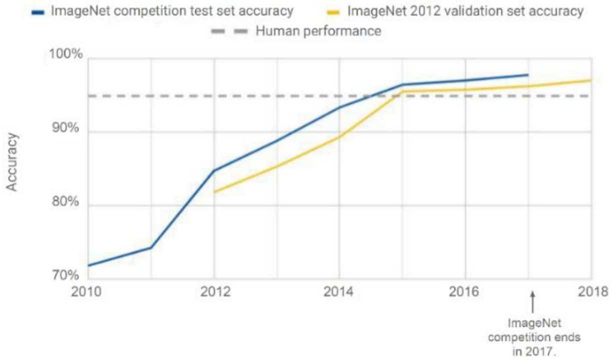
- Cognitive (construct) space:
  - usually non-observable. Performance should correlate with them:
    - agents with a high capabilities profile in this space will imply success for problems with lower difficulty levels in these capabilities.



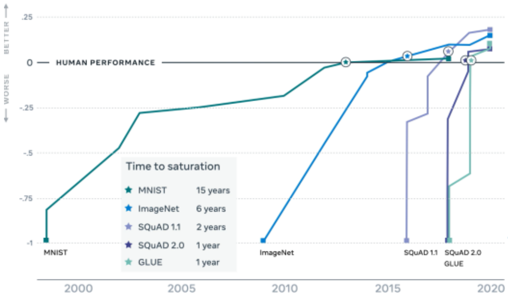
Metrics and Measurement

- Metrics:
  - Capabilities should have a proper scale.
  - Aggregations are not additive from results.
    - More detailed results, annotated instances!
    - No more aggregated results only, please!
    - No more “superhuman” claims, please!
- Measurement:
  - Cover the capabilities space, not the original ***p***.
  - Avoid “challenge-solve-and-replace” dynamics.
  - Explore instance variation:
    - Adaptive testing
    - Adversarial testing

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



AI benchmark saturation over time



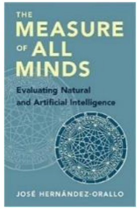
CIFAR10 → CIFAR100,
SQuAD1.1 → SQuAD2.0,
GLUE → SUPERGLUE,
Starcraft → Starcraft II

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

THANKS!

Other Sources and Initiatives:

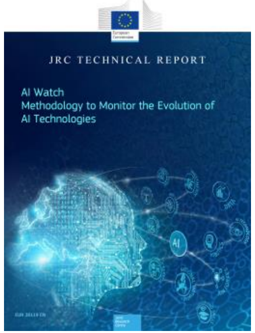
- Other Talks (http://josephorallo.webs.upv.es/)
  - Diversity Unites Intelligence: Measuring Generality
  - Measuring A(G)I Right: Some Theoretical and Practical Considerations
  - Natural and Artificial Intelligence: Measures, Maps and Taxonomies
- Book (http://allminds.org):
  - The Measure of All Minds: Evaluating Natural and Artificial Intelligence, Cambridge University Press 2017



THE MEASURE OF ALL MINDS

Evaluating Natural and Artificial Intelligence

JOSÉ HERNÁNDEZ-ORALLO

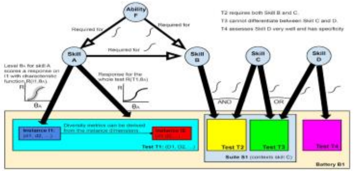


JRC TECHNICAL REPORT

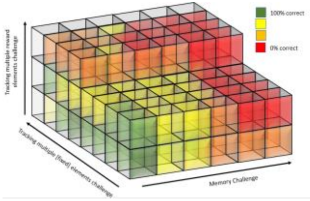
AI Watch

Methodology to Monitor the Evolution of AI Technologies

- The AI Collaboratory: http://aicollaboratory.org/
  - Part of the European Commission's AI watch:
    - https://ec.europa.eu/knowledge4policy/ai-watch_en
- ReCOG-AI and the animal AI environment:
  - Part of the Kinds of Intelligence Programme at the CFI in Cambridge
    - http://lcfi.ac.uk/projects/kinds-of-intelligence



Ability F
Required for
Required for
Required for
Skill A
Skill B
Skill C
Skill D
T2 requires both Skill B and C.
T3 cannot differentiate between Skill C and D.
T4 assesses Skill D very well and has specificity
Level θA for skill A scores a response on I1 with characteristic function R(I1,θA)
Response for the whole test R(T1,θA)
R
θA
R
θA
AND
OR
Instance I1: (d1, d2, ...)
Diversity metrics can be derived from the instance dimensions
Instance I2
Test T1: (D1, D2, ...)
Test T2
Test T3
Test T4
Suite S1 (contexts skill C)
Battery B1



100% correct
0% correct
Tracking multiple reward elements challenge
Tracking multiple (fixed) elements challenge
Memory Challenge