ARTIFICIAL INTELLIGENCE EVALUATION:

Past, present and future*

José Hernández-Orallo
Dep. de Sistemes Informàtics i Computació,
Universitat Politècnica de València
jorallo@dsic.upv.es

Escuela de Verano de Inteligencia Artificial
A Coruña, 3-5 September 2014

* A paper version of this presentation, including full coverage of topics and references, can be found at: http://arxiv.org/abs/1408.6908

WHAT ARE WE AIMING AT?

Image from wikicommons

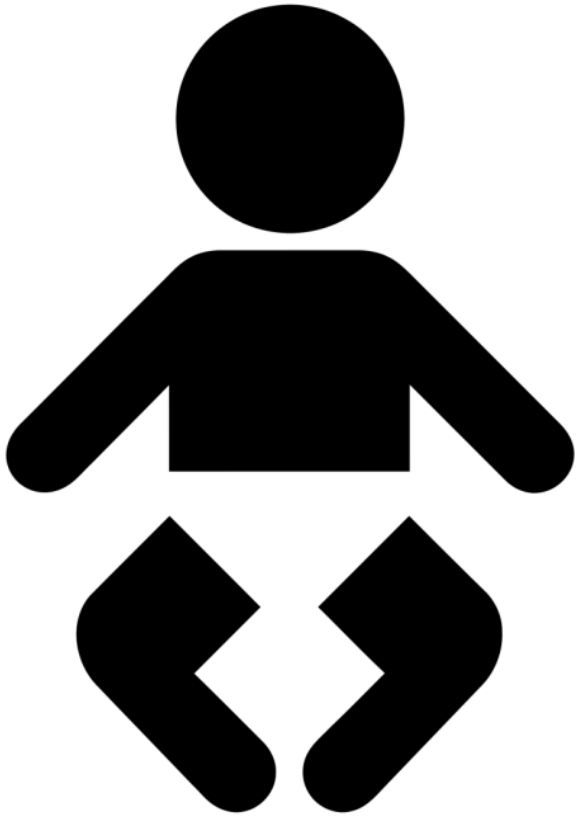
WHAT ARE WE AIMING AT?

Task-oriented

(clear functionality)

Image from wikicommons

WHAT ARE WE AIMING AT?



Warning!
Completely useless until grown up.

ability-oriented

(no functionality)

Image from wikicommons

WHAT ARE WE AIMING AT?

- *A more ambitious view of AI:*

> *"[Artificial Intelligence (AI) is] the science and engineering of making intelligent machines." —John McCarthy (2007)*

- *A more pragmatic view of AI:*

> *"[AI is] the science of making machines do things that would require intelligence if done by [humans]." —Marvin Minsky (1968).*

  - Machines need not be intelligent!
  - They can do the "things" (tasks) **without featuring intelligence**.
    - Once the task is solved, it is no longer an AI problem ("**AI effect**")

OUTLINE

- Why is measuring important for AI?
- PART I. Task-oriented evaluation
  - Types of performance measurement in AI
  - Human discrimination
  - Problem benchmarks
  - Peer confrontation
- PART II. Towards ability-based evaluation
  - What is an ability?
  - The anthropocentric approach: psychometrics
  - The information-theoretic approach
  - Universal psychometrics
- Conclusions

WHY IS MEASURING IMPORTANT?

- *Why* is ***measuring*** important for AI?
  - Measuring and evaluation: at the roots of science and engineering.
  - Disciplines progress when they have *objective* evaluation tools to:
    - Measure the elements and objects of study.
    - Assess the prototypes and artefacts which are being built.
    - Assess the discipline as a whole.
  - E.g., the usual comparison of AI with aeronautics (see, e.g., Russell and Norvig 2009).
    - Aeronautics deals with the construction of flying devices.
      - Measures: mass, speed, altitude, time, consumption, load, wingspan, etc.
      - "Flying" can be defined and evaluated in terms of the above measures.
      - Different specialised devices can be developed by setting different requirements over these measures:
        - Supersonic aircrafts, ultra-light aircrafts, cargo aircrafts, ...

PART I:
TASK-ORIENTED EVALUATION

TASK-ORIENTED EVALUATION

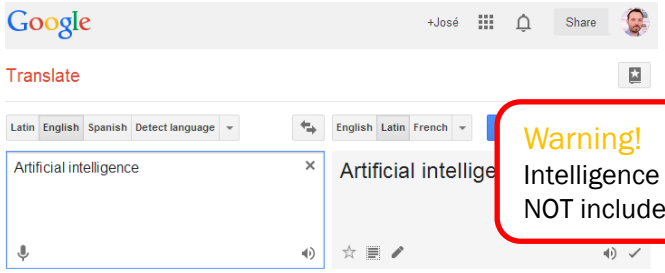
- Specific (task-oriented) AI systems

Warning!
Intelligence
NOT included.

Computer vision, speech recognition, etc.)

Warning!
Intelligence
NOT included.

Prediction and estimation



Warning!
Intelligence
NOT included.

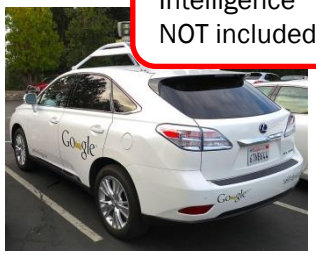
Machine translation, information retrieval, summarisation

Warning!
Intelligence
NOT included.

Robotic navigation

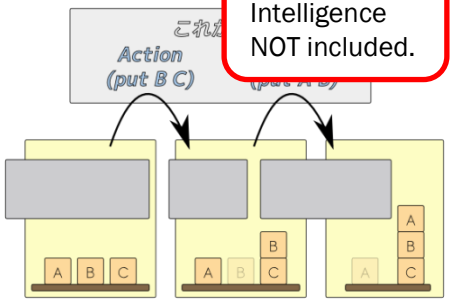
Warning!
Intelligence
NOT included.

Expert systems



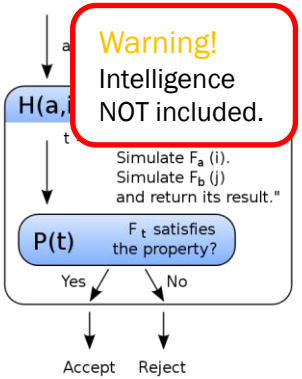
Warning!
Intelligence
NOT included.

Driverless vehicles



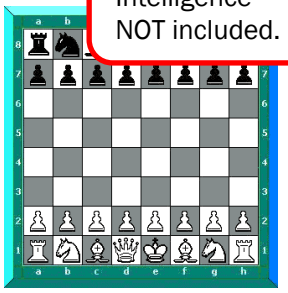
Warning!
Intelligence
NOT included.

Planning and scheduling



Warning!
Intelligence
NOT included.

Automated deduction



Warning!
Intelligence
NOT included.

Game playing

All images from wikicommons

TASK-ORIENTED EVALUATION

- What *instruments* do we have today to evaluate all of them?
  - Application-specific (task-oriented).
    - Linked to a notion of performance for the task (narrow AI).
    - Intelligence is not measured.
    - Best systems usually solve problems in a way that is different to the way humans solve the same problem.
    - Systems include a lot of built-in programming and knowledge for the task.
    - Relatively well-evaluated but with many different (ad-hoc) approaches.

TYPES OF PERFORMANCE MEASUREMENT IN AI

- Consider:
  - A set of problems, tasks or exercises, M.
  - For each exercise $\mu \in M$, we can get a measurement $R(\pi, \mu)$ of the performance of system π.
    - We will use $E[R(\pi, \mu)]$ when the system, the problem or the measurement is non-deterministic and/or imperfect.

- Three common types of aggregated performance metrics:
  - Worst-case performance:
    - $\Phi_{min}(\pi, M) = \min_{\mu \in M} E[R(\pi, \mu)]$
  - Best-case performance:
    - $\Phi_{max}(\pi, M) = \max_{\mu \in M} E[R(\pi, \mu)]$
  - Average-case performance:
    - $\Phi(\pi, M, p) = \sum_{\mu \in M} p(\mu) \cdot E[R(\pi, \mu)]$
      - where $p(\mu)$ is a probability distribution on M.

TYPES OF PERFORMANCE MEASUREMENT IN AI

- Types of white-box (program inspection) assessment.
  - Correct solvers:
    - Performance is defined in terms of time and/or space resources.
    - Classical computational complexity theory.
      - Some AI problems have been analysed in this way.
      - However, it is unreasonable to expect correctness for many AI problems.
  - Approximate solvers:
    - The error of the solution is added to the performance metric.
    - Some other things can be relaxed (e.g., *Probably Approximately Correct*).
  - Game playing and game theory:
    - Several things can be estimated (states, movements, payoff, equilibria).
    - Some games have been solved
      - noughts and crosses (strong), English draughts (weak, J. Schaeffer).
    - Strategies can be compared, optimal strategies can be determined.

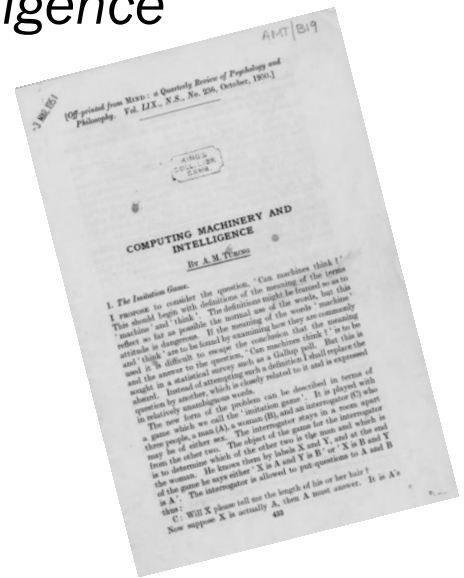
As AI systems become more sophisticated, white-box assessment becomes more difficult, if not impossible (unpredictability of complex systems, like SW).

TYPES OF PERFORMANCE MEASUREMENT IN AI

- Types of black-box (system behaviour) assessment.
  - Human discrimination (observation, scrutiny and/or interview):
    - Assessment is made by and/or against humans. Usually informal.
    - Common in psychology, ethology and comparative psychology.
    - Not usual in AI (except for the Turing Test and variants).
  - Problem benchmarks :
    - Collections or repositories (a set of problems *M* is set up).
      - Common in AI: repositories, problem libraries, corpora, etc.
      - Also usual in (comparative) psychology (e.g., cognitive tests).
    - Problem generators (a class of problems is derived with a generator).
      - This actually defines *M* and *p*.
      - Better characterisation of each problem (e.g., difficulty).
  - Peer confrontation (1-vs-1 or n-vs-n).
    - Evaluates performance in (multi-agent) games from a set of matches.
    - The result is relative to the other participants.
    - Sophisticated performance metrics (e.g., the Elo system in chess).

HUMAN DISCRIMINATION

- *Turing 1950: "Computing Machinery and Intelligence"*
  - A response to nine objections of machine intelligence.
  - The "imitation game" was introduced as a philosophical instrument to help in this response.
  - The game has been (mis-)understood as an actual test, with the standard interpretation:
    - A machine (A), a human (B), and a human interrogator


COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game.

I propose to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's

- Materialisations:
  - Loebner Prize: held since 1991
  - University of Reading 2014 event at the Royal Society.
    - Some interpretations of results stain the reputation of the Turing Test.

HUMAN DISCRIMINATION

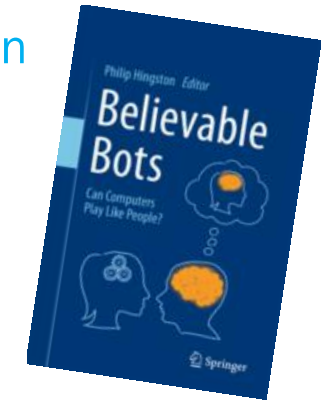
- Is the imitation game a valid test?
  - It has many problems as an intelligence test:
    - It is a test of humanity, **relative** to human characteristics.
    - It is anthropocentric.
    - Neither gradual nor factorial.
    - Needs human intervention (it can't be automated).
    - It takes too much time.
    - Not a sufficient condition.
    - Not a necessary condition.

- Turing is not to be blamed!
  - Not actually conceived by Turing to be a practical test to measure intelligence up to and beyond human intelligence.
  - A great impact in the philosophy and understanding of machine intelligence, but a *negative* impact on its measurement.

HUMAN DISCRIMINATION

- Enhanced Turing Tests:
  - Total Turing Tests, Visual Turing Tests, ...:
    - including sensory information, robotic interfaces, virtual worlds, etc.

- Some other Turing Test variants are more useful.
  - Chatterbot evaluation.
    - Applications: personal assistants, games, ...
  - Avatar evaluation:
    - Videogames.
      - Bots can fool opponents into thinking it is another human player
  - Interesting new notions:
    - Bots have to be **believable** (Hingston 2012).
    - Bots have to be enjoyable, fun, etc.



HUMAN DISCRIMINATION

- Example: BotPrize (http://botprize.org/)
  - Held on 2008, 2009, 2010, 2011, 2012, 2014 (Spain!)

  - Rules:
    - Uses the "DeathMatch game type for the First-Person Shooter, Unreal Tournament 2004".
      - The bots don't process the image but receive a description of it through textual messages in a language through the GameBots2004 interface (Pogamut).
    - Chatting is disabled (it's not a chatbot competition)
    - The player that looks most "human" wins the game.
    - There is a "judging gun". Bots also judge.
    - The judges play, trying to play normally (a prize for the judges exists for those that are considered more "human" by other judges).

HUMAN DISCRIMINATION

- Example: BotPrize. Improvements.
  - "Believability" is said to be better assessed from a *third-person perspective* (judging recorded video of other players without playing) than a *first-person perspective* (Togelius et al 2012).
    - Reason: human judges can concentrate on judging and not on not being killed or aiming at high scores.
    - This third-person perspective is included in the 2014 competition using a crowdsourcing platform:
      - (Llargues-Asensio et al. 2014, Expert Systems with Applications)
    - In the 2014 edition there are two judging systems:
      - First-Person Assessment (FPA): BotPrize in-game judging system.
      - Third-Person Assessment (TPA): crowdsourcing platform.
  - Challenges: richer (and more difficult) representation of the environment (such as a graphical processing as in the Arcade Learning Environment).

PROBLEM BENCHMARKS

- *M* is a set of problems.
- The quality of these evaluations depend on *M*.
  - *M* is usually known before the evaluation.
    - On occasions, the solutions are also known beforehand or can be inferred by humans.
    - Most systems actually embed what the researchers have learnt from the problem.

> These benchmarks actually evaluate the researchers, not their systems!

    - Much worse if the selection of *M* is made by the researchers (e.g., selection of datasets from the UCI repository).

PROBLEM BENCHMARKS

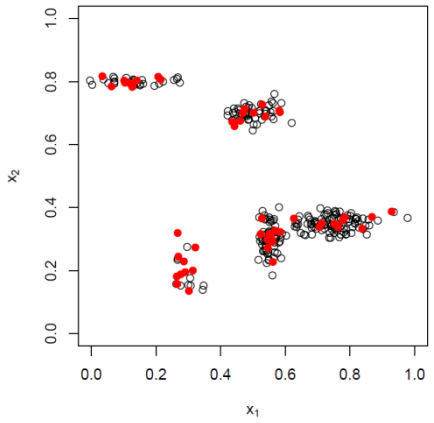
- Much better if *M* is very large or infinite and examples are samples or generated from *M*.
  - It is not always easy to generate a large *M* of realistic problems.
    - Generators can be based on:
      - Some prototypes with parameter variations.
      - Problem representation languages
        - Not easy to rule out unusable problems.
    - A general and elegant approach is to determine a probabilistic or stochastic generator (e.g. a grammar) of problems, which directly defines the probability *p* in the average-case performance formula:
      - $\Phi(\pi, M, p) = \sum_{\mu \in M} p(\mu) \cdot \mathrm{E}[R(\pi, \mu)]$

PROBLEM BENCHMARKS

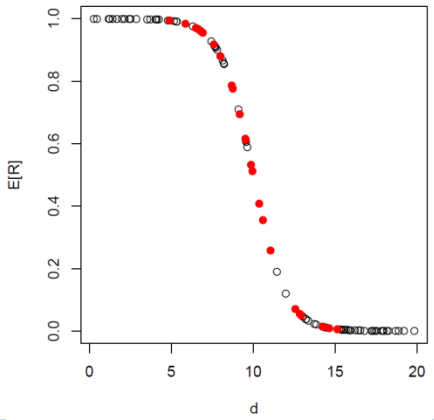
- Distinguish the problem set from an effective evaluation.
  - Finite test: limited number of exercises *n* that we can administer.
    - The goal is to reduce the variance of the measurement given *n*.
  - No-sampling approach:
    - Sort by decreasing *p* and evaluate the system with the first *n* exercises.
      - This maximises the accumulated mass for *p* for a given *n*.
      - It is highly predictable. Systems will specialise on the first *n* exercises.
      - Not very meaningful when *R* is not deterministic and/or not completely reliable. Repeated testing may be needed.
  - Random sampling using *p:*
    - With replacement (as *R* is usually non-deterministic and/or not completely reliable).
    - If *M* and *p* define the benchmark, is probability-proportional sampling on *p* the best way to evaluate systems?
      - No, in general. There are better ways of approximating $\Phi(\pi, M, p)$.

PROBLEM BENCHMARKS

- Information-driven sampling.
  - Related to *importance sampling* and *stratified sampling*. We use a different probability distribution for sampling.
  - Diversity-driven sampling:
    - Given a similarity, a set of features or any other way to determine how similar two exercises are.
    - We need to sample on *M* such that:
      - the accumulated mass on *p* is high.
      - diversity has to be maximised.
  - Difficulty-driven sampling.
    - The idea is to choose a range of difficulties with high weight.
    - Difficulty is defined as function d: $M \rightarrow \Re$.
      - $d(\mu)$ is monotonically decreasing on $E_{\pi \in \Omega}[\Phi(\pi, \mu, p)]$
    - We need to sample on M such that only the informative difficulties are covered.



Covering *p* without sampling very similar exercises repeatedly, and correcting the results accordingly (e.g., cluster sampling)



The results below d=5 and above d=15 can be assumed to be known, so effort is focussed on the relevant range.

PROBLEM BENCHMARKS

- Adaptive sampling
  - Informative-driven sampling can be made adaptive (e.g. adaptive clustering testing, or adaptive difficulty-based testing).
  - In Psychometrics, Computerised Adaptive Testing (CAT) uses difficulty to estimate the value for Φ in very few iterations.
    - Item Response Theory (IRT) describes expected outcome of a population for a given item (exercise) with *Item Response Functions.*
      - *Proficiency (θ) corresponds to difficulty.*

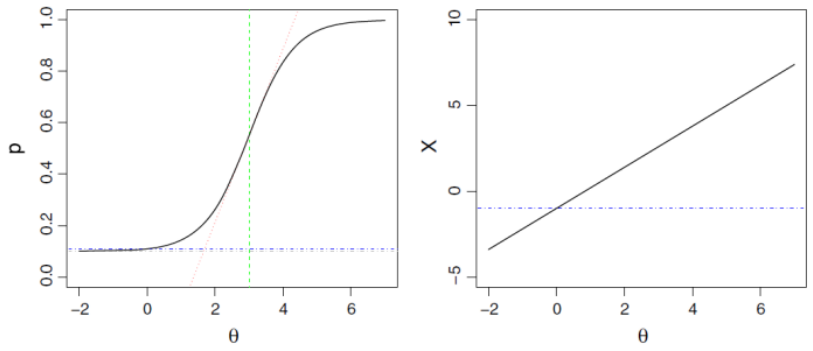


Fig. 1 *Left* Item response function (or *curve*) for a binary score item with the following parameters for the logistic model: discrimination $a = 1.5$, item location $b = 3$, and chance $c = 0.1$. The discrimination is shown by the slope of the curve at the midpoint: $a(1 - c)/4$ (in *dotted red*), the location is given by b (in *dashed green*) and the chance is given by the *horizontal line* at c (in *dashed-dotted grey*), which is very close to the zero-ability expected result $p(\theta) = z$ (here at 0.11). *Right* A linear model for a continuous score item with parameter $z = -1$ and $\lambda = 1.2$. The *dashed-dotted line* shows the zero-ability expected result (color figure online)

```
Item  Theta   SE     -3.......-2.........-1.........0.........+1.........+2.........+3
  0   0.00*  1.00*            ---------------------X----------------------
  1   4.00*  1.00*                                 .                      ---------->
  2   0.11   0.52                          ----------I----------
  3   0.20   0.45                           ----------C----------
  4  -0.04   0.35                          -------I--------
  5   0.05   0.32                            ------C------
  6  -0.13   0.29                          ------I-------
  7  -0.07   0.27                           ------C------
  8  -0.18   0.25                           -----I------
  9  -0.25   0.25                          -----I------
 10  -0.18   0.23                           -----C-----
 11  -0.27   0.23                          -----I-----
 12  -0.21   0.22                           ----C------
 13  -0.26   0.22                           ----I-----
 14  -0.34   0.22                          ----I------
 15  -0.37   0.22                         -----I-----
 16  -0.33   0.20                          ----C-----
 17  -0.29   0.19                           ----C---
 18  -0.33   0.19                          ----I-----
 19  -0.38   0.19                          ----I-----
 20  -0.34   0.18                          ----C-----
 21  -0.30   0.18                           ----C---
 22  -0.27   0.17                           ----C---
 23  -0.29   0.17                           ----I---
 24  -0.26   0.17                            ---C---
 25  -0.28   0.16                           ----I---
 26  -0.30   0.16                           ----I---
 27  -0.27   0.16                            ---C---
 28  -0.25   0.15                            ---C---
 29  -0.23   0.15                            ---C---
 30  -0.21   0.15                            ---C---
```

An example of an IRT-based adaptive test (freely adapted from Fig. 8 in Weiss 2011).

PROBLEM BENCHMARKS

- Example: “The UCI test”
  - UCI (and other machine learning repositories) and Kaggle competitions.
  - Typically referred to as "The UCI test" (Macià & E. Bernardó-Mansilla 2014) or the "de facto approach" (Japkovich - Shah 2011).
  - Follows the general form:
    - $\Phi(\pi, M, p) = \sum_{\mu \in M} p(\mu) \cdot \mathrm{E}[R(\pi, \mu)]$
    - M is the repository, p is the choice of datasets and R is one particular performance metric (accuracy, AUC, Brier score, F-measure, MSE, etc.)
  - "The UCI test" is a bona-fide approaches.
  - Actually mixes of a problem benchmark with peer confrontation:
    - Problem benchmark: there is a repository (M), but only a few problems are cherry-picked (p is changing and arbitrary).
    - Peer confrontation: only a few competitors are cherry-picked without much effort on choosing their best parameters.
      - Algorithms can be compared 1vs1 using statistical tests.
        - Cross-validation or other repetition approaches are used to reduce the variance of $R(\pi, \mu)$ so that we have more “wins”.

PROBLEM BENCHMARKS

- Example: "The UCI test". Improvements
  - UCI+ proposal (Macià & E. Bernardó-Mansilla 2014, Information Sciences).
    - Characterise UCI to provide more diversity
    - Use complexity measures from (Ho & Basu, 2002, TPAMI). What's a "challenging" problem? → "difficulty".
    - Include an artificial dataset generator. It is a distortion-based generator (similar to C. Soares's UCI++).
    - Ideas about sharing results (e.g., openml.org), automated submission, ...
  - Other improvements.
    - Use of complexity measures to derive how representative a problem is of the whole distribution and to sample more adequately.
    - Pattern-based generator instead of distortion-based generators.
      - E.g., try to define p with a stochastic generative grammar.

PEER CONFRONTATIONS

- Matches are played between peers.
  - How can we derive an independent measure of performance?
  - Results are relative to the opponents.
    - We define the set Ω of all the opponents. In a way, the set of problems M is enriched (or even substituted) by one single game (e.g. chess) with different competitors.
    - How to compare results between two different competitions if opponents are different? How to compare progress?
      - If there are common players, we can use rankings, such as the Elo ranking, to see whether there are progress.
    - Systems can specialise to the kind of opponents that are expected in a competition. This is usual in sports.

PEER CONFRONTATIONS

- Games and multi-agent environments could be evaluated against standardised opponents.
  - However, how to choose a standardised opponent?
    - If the opponent is known, the systems can be specialised to the opponent.
      - E.g., checkers players could specialise to play against Chinook.
  - Opponent generators?
    - Random actions → too bad.
    - Use an agent-language for the generation of Ω.
    - How can we assess whether the Ω has sufficiently difficulty and discriminative power?
      - A difficult problem, analysed in (Hernandez-Orallo 2014, JAAMAS).
    - We can give more information and resources to these players to make them more competitive.

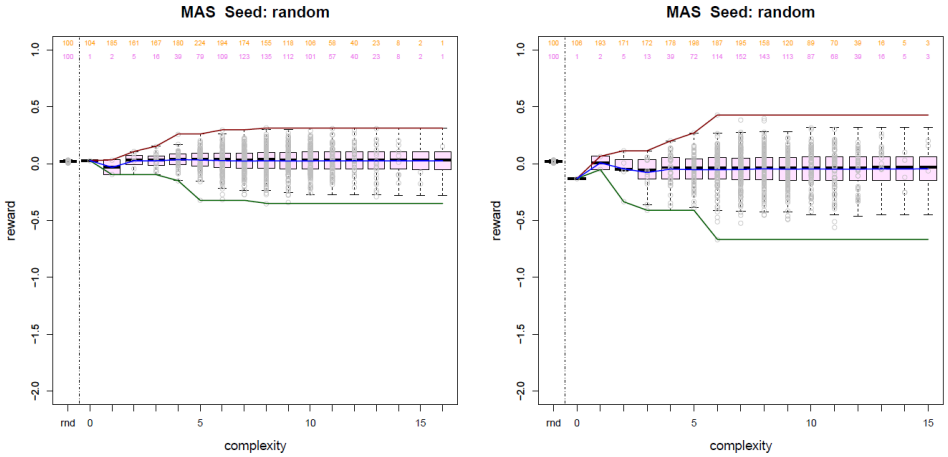


Figure 4: We show the distributions of reward (roughly corresponding to R in this paper) for different configurations for the multi-agent system SCMAS introduced in [17]. Left: the plot show the results when we confront each of the 2,000 policies with 50 different teams of competitors (with different seeds for the generator also). This means that we have $2,000 \times 50 = 100,000$ experiments (300 environment steps each). This is what we see on the bottom-left plot. Right: results when we choose the best 8 agents from the previous experiment. We see a wider range of results (but note that the average reward is lower).

PEER CONFRONTATIONS

- Example: General Game Competition
  - Running yearly: 2005-2014 (http://games.stanford.edu/)
  - Available server and languages.
  - Rules:
    - "General game players are systems able to accept descriptions of arbitrary games at runtime."
    - "They do not know the rules until the games start."
      - Games are described in the language GDL (Game description language). The description of the game is given to the players.
    - "They should be able to play simple games (like Tic Tac Toe) and complex games (like Chess), games in static or dynamic worlds, games with complete and partial information, games with varying numbers of players, with simultaneous or alternating play, with or without communication among the players, and so forth."
      - For the competition, games are chosen (non-randomly, manually by the organisers) from the pool of games already described in GDL and new games can be introduced for the competition.
        - Game specialisation is difficult.

PEER CONFRONTATIONS

- Example: General Game Competition: improvements
  - A more sophisticated analysis of how difficult and representative games are.
  - Derivation of rankings and the accumulation of former participants for the following competitions.
  - Learning without the description of the game, as a reinforcement learning problem (where the system learns the rules from many matches) could be interesting:
    - "Integration of General Game Playing with RL-glue" (http://users.dsic.upv.es/~flip/RLGGP/ggp-integration.pdf)
    - Like the reinforcement learning competition but without a set of predefined problems.

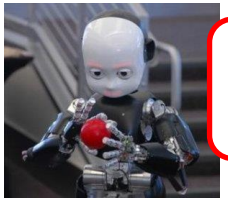
EXAMPLES OF EVALUATION SETTINGS

- Specific domain evaluation settings:
  - CADE ATP System Competition → PROBLEM BENCHMARKS
  - Termination Competition → PROBLEM BENCHMARKS
  - The reinforcement learning competition → PROBLEM BENCHMARKS
  - Program synthesis (Syntax-guided synthesis) → PROBLEM BENCHMARKS
  - Loebner Prize → HUMAN DISCRIMINATION
  - Robocup and FIRA (robot football/soccer) → PEER CONFRONTATION
  - International Aerial Robotics Competition (pilotless aircraft) → PROBLEM BENCHMARKS
  - DARPA driverless cars, Cyber Grand Challenge, Rescue Robotics → PROBLEM BENCHMARKS
  - The planning competition → PROBLEM BENCHMARKS
  - General game playing AAAI competition → PEER CONFRONTATION
  - BotPrize (videogame player) contest (2014 in Spain) → HUMAN DISCRIMINATION
  - World Computer Chess Championship → PEER CONFRONTATION
  - Computer Olympiad → PEER CONFRONTATION
  - Annual Computer Poker Competition → PEER CONFRONTATION
  - Trading agent competition → PEER CONFRONTATION
  - Robo Chat Challenge → HUMAN DISCRIMINATION
  - UCI repository, PRTools, or KEEL dataset repository. → PROBLEM BENCHMARKS
  - KDD-cup challenges and ML kaggle competitions → PROBLEM BENCHMARKS
  - Machine translation corpora: Europarl, SE times corpus, the euromatrix, Tenjinno competitions… → PROBLEM BENCHMARKS
  - NLP corpora: linguistic data consortium, … → PROBLEM BENCHMARKS
  - Warlight AI Challenge → PEER CONFRONTATION
  - The Arcade Learning Environment → PROBLEM BENCHMARKS
  - Pathfinding benchmarks (gridworld domains) → PROBLEM BENCHMARKS
  - Genetic programming benchmarks → PROBLEM BENCHMARKS
  - CAPTCHAs → HUMAN DISCRIMINATION
  - Graphics Turing Test → HUMAN DISCRIMINATION
  - FIRA HuroCup humanoid robot competitions → PROBLEM BENCHMARKS
  - …

PART II: TOWARDS ABILITY-ORIENTED EVALUATION

TOWARDS ABILITY-ORIENTED EVALUATION

- How can we evaluate more general AI systems?



Warning!
Some intelligence
MAY BE included.

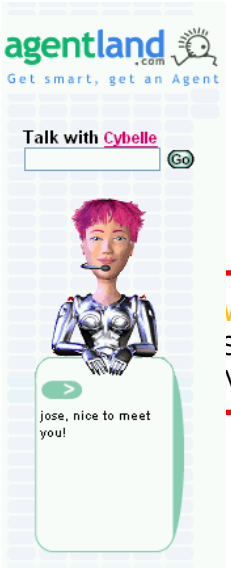
Cognitive robots

Warning!
Some intelligence
MAY BE included.

Pets, animats and other artificial companions

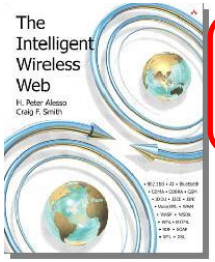
Warning!
Some intelligence
MAY BE included.

Smart buildings



Warning!
Some intelligence
MAY BE included.

Agents, avatars, chatbots



Warning!
Some intelligence
MAY BE included.

Web-bots, Smartbots, Security bots…

Warning!
Some intelligence
MAY BE included.

Intelligent assistants

TOWARDS ABILITY-ORIENTED EVALUATION

- Artificial Intelligence: gradually catching up (and then outperforming) humans' performance for more and more tasks:
  - Calculation: XVIIIth and XIXth centuries
  - Cryptography: 1930s-1950s
  - Simple games (noughts and crosses, connect four, …): 1960s
  - More complex games (draughts, bridge): 1970s-1980s
  - Printed (non-distorted) character recognition: 1970s
  - Data analysis, statistical inference, 1990s
  - Chess (Deep Blue vs Kasparov): 1997
  - Speech recognition: 2000s (in idealistic conditions)
  - TV Quiz (Watson in Jeopardy!): 2011
  - Driving a car: 2010s
  - Texas hold 'em poker: 2010s
  - Translation: 2010s (technical documents)
  - …

TOWARDS ABILITY-ORIENTED EVALUATION

- Tasks are classified as: (Rajani 2010, Information Technology)
  - *optimal*: it is not possible to perform better
  - *strong super-human*: performs better than all humans
  - *super-human*: performs better than most humans
  - *par-human*: performs similarly to most humans
  - *sub-human*: performs worse than most humans
- This view of "progress in artificial intelligence" is misleading.
  - All these systems are task-oriented systems.

No AI system can do (*or can learn to do*) **all** these things!

*Despite pitiful **big-switch** approaches*

But this system can:

A different perspective for AI evaluation:
"machines do tasks they have never seen and have not been prepared for beforehand."

WHAT IS AN ABILITY?

- We are talking about *cognitive* abilities:

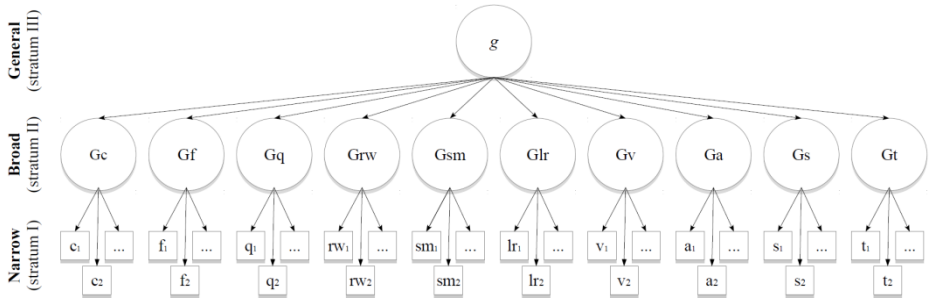
> A cognitive ability is a property of individuals which allows them to perform well in a *range* of information-processing tasks.

  - The ability is *required*.
    - Performance is much worse ***without featuring the ability***.
    - Note that the ability is *necessary* but it does *not* have to be *sufficient*.
      - E.g., spatial abilities are necessary but not sufficient for driving a car.
  - *General*, covering a range of tasks.
  - Problem: abilities have to be conceptualised and identified.
    - Abilities are constructs while tasks are instruments.

Figure adaptation courtesy of Fernando Martínez-Plumed

WHAT IS AN ABILITY?

- Many arrangements of cognitive abilities have been identified.
  - For instance, the Cattell-Horn-Carroll theory:
    - Broad abilities:
      - Crystallised Intelligence (Gc), Fluid Intelligence (Gf), Quantitative Reasoning (Gq), Reading and Writing Ability (Grw), Short-Term Memory (Gsm), Long-Term Storage and Retrieval (Glr), Visual Processing (Gv), Auditory Processing (Ga), Processing Speed (Gs) and Decision/Reaction Time/Speed (Gt)

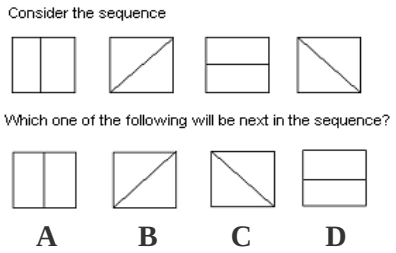


  - The broad abilities seem to correspond to subfields in AI:
    - problem solving, use of knowledge, reasoning, learning, perception, natural language processing, … (from Russell and Norvig 2009).

Figure adaptation courtesy of Fernando Martínez-Plumed

THE ANTHROPOCENTRIC APPROACH: PSYCHOMETRICS

- Goal: evaluate the intellectual abilities of human beings
  - Developed by Binet, Spearman and many others at the end of the XIXth century and first half of the XXth century.
    - Culture-fair: no “idiots savants”.
  - A joint index is usually determined, known as IQ (Intelligence Quotient).
    - **Relative** to a population: initially normalised against the age, then normalised (μ=100, σ=15) against the adult average.
- IQ tests are easy to administer, fast and accurate.
  - Used by companies and governments, essential in education and pedagogy.
  - Tests are factorised.
    - g factor (general intelligence),
    - verbal comprehension,
    - spatial abilities,
    - memory,
    - inductive abilities,
    - calculation and deductive abilities

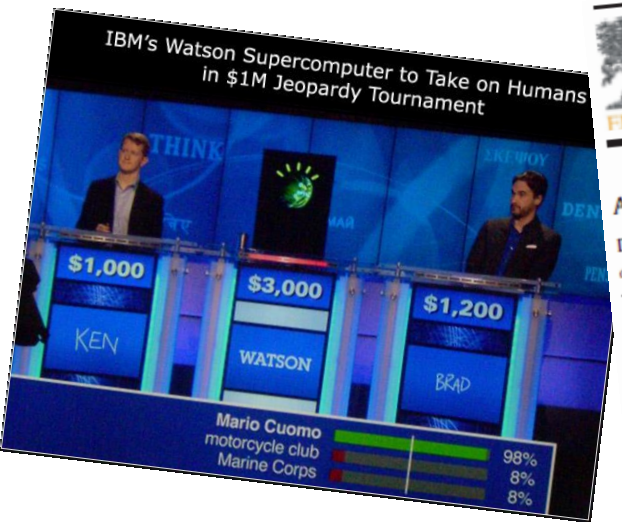


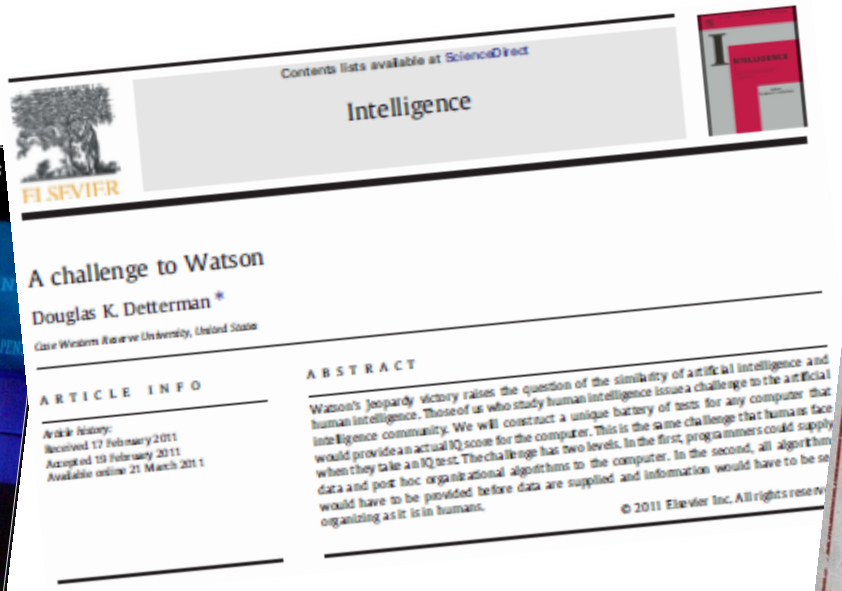
Complete the matrix

2	4	8
3	6	12
4	8	?

THE ANTHROPOCENTRIC APPROACH: PSYCHOMETRICS

- Let's use them for machines!
  - This has been suggested several times in the past.
- Detterman, editor of the *Intelligence Journal*, made this suggestion serious and explicit: "A challenge to Watson (2011)"
  - As a response to specific domain tests and landmarks (such as Watson).




Contents lists available at ScienceDirect

Intelligence

ELSEVIER

A challenge to Watson

Douglas K. Detterman *

Case Western Reserve University, United States

ARTICLE INFO

Article history:
Received 17 February 2011
Accepted 19 February 2011
Available online 21 March 2011

ABSTRACT

Watson's Jeopardy victory raises the question of the similarity of artificial intelligence and human intelligence. Those of us who study human intelligence issue a challenge to the artificial intelligence community. We will construct a unique battery of tests for any computer that would provide an actual IQ score for the computer. This is the same challenge that humans face when they take an IQ test. The challenge has two levels. In the first, programmers could supply data and post hoc organizational algorithms to the computer. In the second, all algorithms would have to be provided before data are supplied and information would have to be se organizing as it is in humans.

© 2011 Elsevier Inc. All rights reserv

THE ANTHROPOCENTRIC APPROACH: PSYCHOMETRICS

- Hold on!
  - In 2003, Sanghi & Dowe implemented a small program (in Perl) which could score relatively well on many IQ tests.
    - A 3rd year student project
    - Less than 1000 lines of code
      - (a big-switch approach)

This made the point unequivocally: this program is not intelligent

Warning!
Intelligence
NOT included.

Test	I.Q. Score	Human Average
A.C.E. I.Q. Test	108	100
Eysenck Test 1	107.5	90-110
Eysenck Test 2	107.5	90-110
Eysenck Test 3	101	90-110
Eysenck Test 4	103.25	90-110
Eysenck Test 5	107.5	90-110
Eysenck Test 6	95	90-110
Eysenck Test 7	112.5	90-110
Eysenck Test 8	110	90-110
I.Q. Test Labs	59	80-120
Testedich.de:I.Q. Test	84	100
I.Q. Test from Norway	60	100
Average	96.27	92-108

THE ANTHROPOCENTRIC APPROACH: PSYCHOMETRICS

- Response to Detterman:
  - "IQ tests are not for machines, yet" (Dowe & Hernandez-Orallo 2012, Intelligence Journal)
  - IQ tests take many things for granted:
    - They are anthropocentric.
      - On top of that, they are specialised to the average human.
      - Tests are broader when evaluating small children, people with disabilities, etc.?
  - Can we devise different IQ test batteries such that AI systems (e.g., Sanghi and Dowe's program) fail?
    - This would end up as a psychometric CAPTCHA.
  - IQ tests are increasingly more used in AI
    - For a survey, Hernandez-Orallo et al. 2015, AIJ.

THE ~~ANTHROPOCENTRIC~~ CHIMPOCENTRIC APPROACH!

- Animal evaluation and comparative psychology
  - Animals and compared (abilities are "**relative to…**")
- Is it isolated from psychometrics?
  - Partly it was, but it is becoming closer and closer, especially when comparing apes and human children
- Applicable to machines?
  - Not directly.
  - But many ideas (and the overall perspective) are useful:
    - Use of rewards and interfaces
    - Abilities as concepts and tests as instruments.
    - Testing social abilities (co-operation and competition) is common.
    - No prejudices.
    - Non-anthropocentric:
      - exploring the animal kingdom.
      - humans as a special case.

Images from BBC One documentary: "Super-smart animal": http://www.bbc.co.uk/programmes/b01by613

THE INFORMATION-THEORETIC APPROACH

- A different approach to evaluation started in the late 1990s
  - Algorithmic Information Theory (Turing, Shannon, Solmonoff, Kolmogorov, Chaitin, Wallace)
    - Kolmogorov complexity, $K_U(s)$: shortest program for machine U which describes/outputs an object s (e.g., a binary string).
    - Algorithmic probability (universal distribution), $p_U(s)$: the probability of objects as outputs of a UTM U fed by 0/1 from a fair coin.
      - Immune to the NFL theorem (every computable distribution can be approximated by a universal distribution).
    - Both are related (under prefix-free or monotone TMs): $p_U(s) = 2^{-K_U(s)}$
    - Invariance theorem: the value of $K(s)$ (and hence $p(s)$) for two different reference UTMs U_1 *and* U_2 only differs by (at most) a constant (which is independent of s).
    - $K(s)$ is incomputable, but approximations exist (Levin's Kt).
  - Formalisation of Occam's razor: shorter is better!
  - Compression and inductive inference (and *learning*): two sides of the same coin (Solomonoff, MML, ...).

THE INFORMATION-THEORETIC APPROACH

- Compression and intelligence
  - Compression-enhanced Turing Tests (Dowe & Hajek 1997-1998).
    - A Turing Test which includes compression problems.
    - By ensuring that the subject needs to compress information, we can make the Turing Test more sufficient as a test of intelligence and discard objections such as Searle's Chinese room.
  - But it is still a Turing Test…

THE INFORMATION-THEORETIC APPROACH

- Intelligence *definition* and *test* (C-test) based on algorithmic information theory (Hernandez-Orallo 1998-2000).
  - Series are generated from a TM with a general alphabet and some properties (projectibility, stability, ...).

$k = 9$: a, d, g, j, . . . Answer : m
$k = 12$: a, a, z, c, y, e, x, . . . Answer : g
$k = 14$: c, a, b, d, b, c, c, e, c, d, . . . Answer : d

- Intelligence is the result of a test:

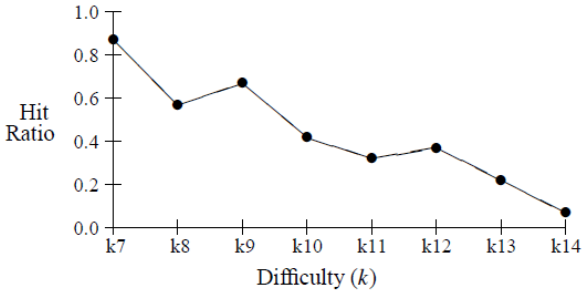
$$I_U(\sigma) \triangleq \sum_{e \in \mathbb{E}} \rho_U(e) \cdot w_U(e, \varepsilon) \cdot H_e^\sigma \approx \frac{1}{n'} \sum_{k=K_{min} \ldots K_{max}} k^\varepsilon \cdot \sum_{e_{j,k} \text{ with } Kt_U(e_j)=k, j=1 \ldots n} H_{e_{j,k}}^\sigma$$

where U is the reference machine, σ is the subject, $\mathbb{E}$ is the set of all possible exercises (stable projectible series), ε is the weight depending on the difficulty, H_e^σ is whether the subject σ makes the exercise e correctly, and n' is a normalisation factor which depends on n, ε, K_{max} and K_{min}.

  - Bears similarities with our aggregated measures using *M* and *p*.

THE INFORMATION-THEORETIC APPROACH

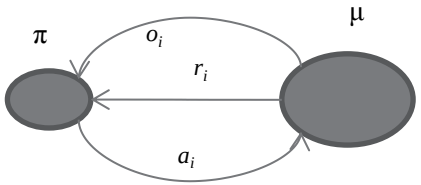
- Very much like IQ tests, but formal and well-grounded :
  - exercises are not chosen arbitrarily.
  - the right solution (projection of the sequence) is 'unquestionable'.
  - Item difficulty derived in an 'absolute' way.
    - *Human performance correlated with the absolute difficulty (k) of each exercise and IQ tests for the same subjects:*



- This is IQ-test re-engineering!
  - However, some simple programs can ace on them (e.g., Sanghi and Dowe 2003).
  - They are static (series): no planning/"action" required.
  - Only covers general intelligence. Other abilities (Hernández-Orallo 2000b, NIST)

THE INFORMATION-THEORETIC APPROACH

- Intelligence as performance in a range of worlds. (Hutter 2000, Dobrev 2000, 2005)
  - Worlds: interactive environments
    - R is understood as the degree of success
  - The set of worlds *M* is described by Turing machines.
    - Bounded or weighted by Kolmogorov complexity.



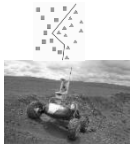
  - Intelligence is measured as an average, following the average-case evaluation:
    - $\Phi(\pi, M, p) = \sum_{\mu \in M} p(\mu) \cdot \mathrm{E}[R(\pi, \mu)]$
  - *"Universal Intelligence"* (Legg and Hutter 2007): much better formalised.
  - Both are interactive extensions of C-tests from sequences to environments...

- Problems:
  - For both approaches, the mass of the probability measure goes to a few environments.
  - *M* or the probability distribution is not computable.
  - Most environments are not really discriminative (Dobrev discusses this issue briefly).
  - (Legg and Hutter) There are two infinite sums (environments and interactions).
  - Time/speed is not considered for the environment or for the agent.

UNIVERSAL PSYCHOMETRICS

- A snapshot of the fragmentation of intelligence evaluation…

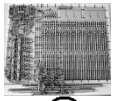
- Human-discriminative (e.g., Turing test):
  1. Held in a human natural language.
  2. The examinees 'know' it is a test.
  3. Interactive.
  4. Adaptive.
  5. Relative to humans.



- Problem benchmarks:
  1. Task-specific tests.
  2. Choice of problems not always representative.
  3. Generally non-adaptive.
  4. Risk of problem overfitting.

- Peer confrontation:
  1. Task-specific tests.
  2. Highly dependent on the opponents (relative to a population)
  3. Standard measurements difficult to obtain
  4. A good match arrangement necessary for reliability of results.

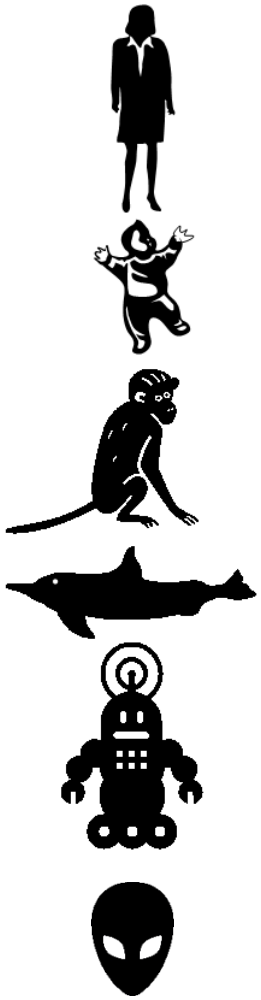
- IQ tests:
  1. Human-specific tests.
  2. The examinees know it is a test.
  3. Generally non-interactive.
  4. Generally non-adaptive (pre-designed set of exercises)
  5. Relative to a population



- Tests and definitions based on AIT
  1. Interaction highly simplified.
  2. The examinees do not know it is a test. Rewards may be used.
  3. Sequential or interactive.
  4. Non-adaptive.
  5. Formal foundations.

- Animal (and children) intelligence evaluation:
  1. Perception and action abilities assumed.
  2. The examinees do not know it is a test. Rewards are used.
  3. Interactive.
  4. Generally non-adaptive.
  5. Comparative (relative to other species).

UNIVERSAL PSYCHOMETRICS



- Can we construct tests for all of them?
  - Without knowledge about the examinee,
  - Derived from computational principles,
  - Non-biased (species, culture, language, etc.)
  - No human intervention,
  - Producing a score,
  - Meaningful,
  - Practical, and
  - Anytime.

UNIVERSAL PSYCHOMETRICS

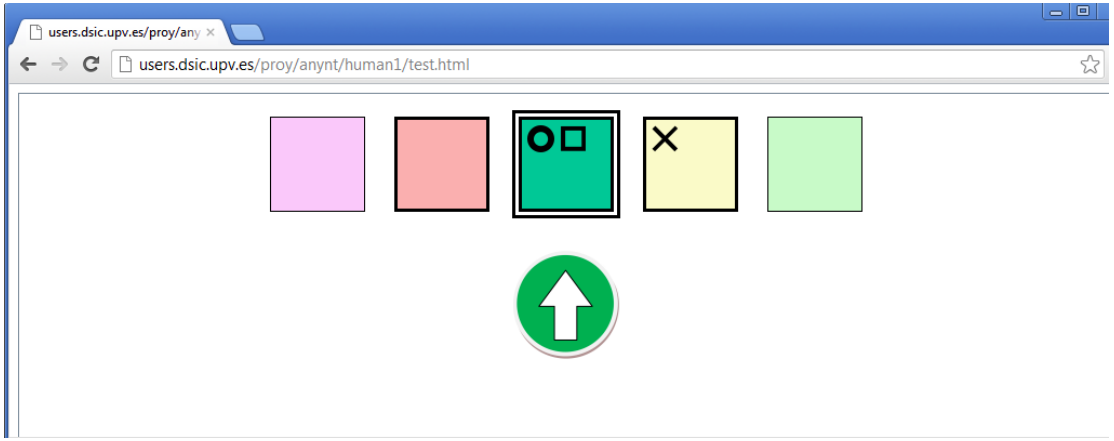
- Anytime universal test (Hernandez-Orallo & Dowe 2010, Artificial Intelligence):
  - The class of environments is carefully selected to be discriminative.
  - Environments are randomly sampled from that class.
    - Starts with very simple environments.
    - Complexity of the environments adapts to the subject's performance.
  - The speed of interaction adapts to the subject's performance.
  - Includes time.
  - It can be stopped anytime.

UNIVERSAL PSYCHOMETRICS

- The anYnt project (2009-2011):

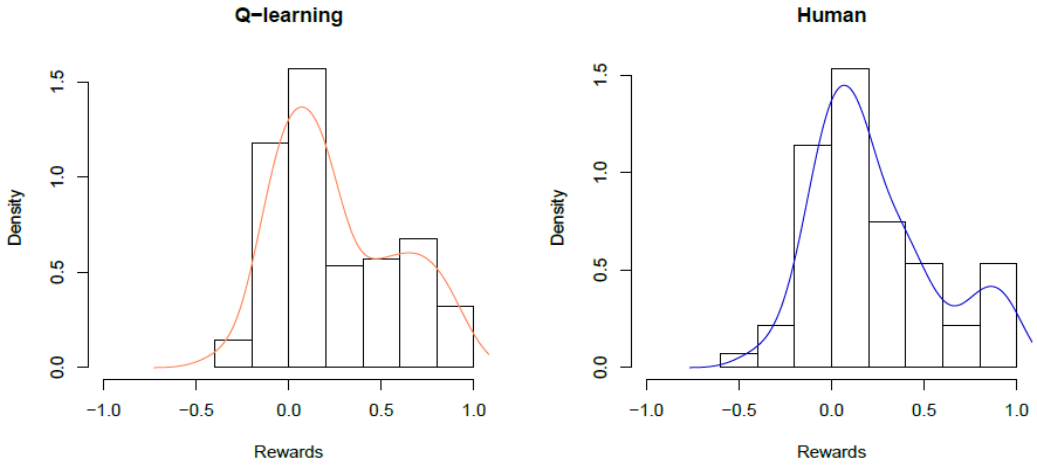
  http://users.dsic.upv.es/proy/anynt/

  - Goal: evaluate the feasibility of a universal test.
    - What do environments look like?
      - An environment class Λ was devised.
    - The complexity/difficulty function Kt^{max} was chosen.
    - An interface for humans was designed.



UNIVERSAL PSYCHOMETRICS

- Experiments (2010-2011):
  - The test is applied to humans and an AI algorithm (Q-learning):



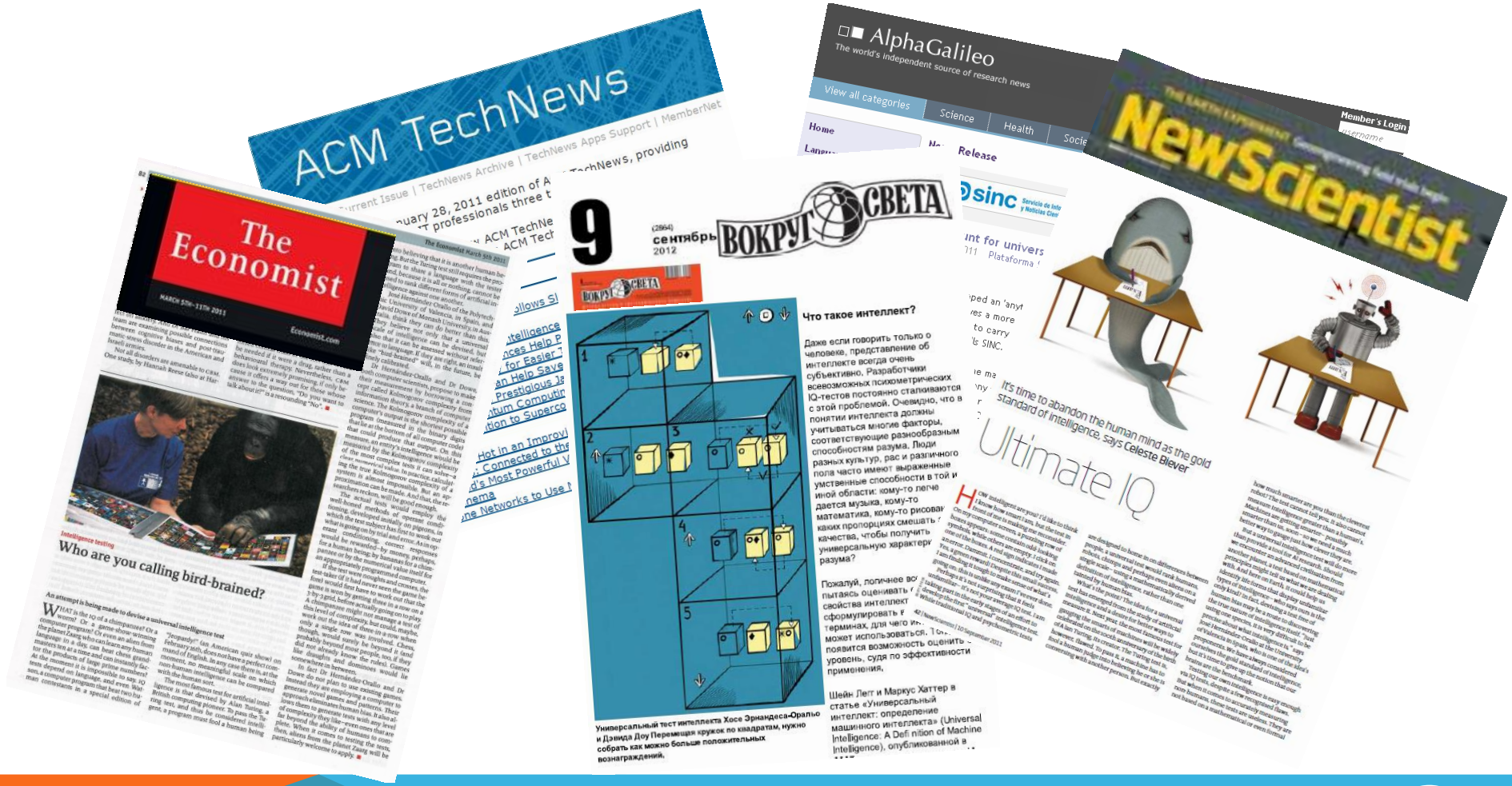
  - Impressions:
    - The test is useful to compare and scale systems of the same type.
    - The results do not reflect the actual differences between humans and Q-learning.

UNIVERSAL PSYCHOMETRICS

- How should this *Popperian refutation* be interpreted?
  - It was a prototype: many simplifications made.
  - It is not adaptive (not anytime)
  - Absence of noise: specially beneficial for AI agents.
  - Patterns have low complexity.
  - The environment class may be richer.
  - More factors may be needed.
  - No incremental knowledge acquisition.
  - No social behaviour (environments weren't multi-agent).
- Are universal tests impossible?
  - All the above issues should be explored before dismissing this idea.

UNIVERSAL PSYCHOMETRICS

- anYnt project media coverage! (despite the limited results)



UNIVERSAL PSYCHOMETRICS

- *Something went very wrong here…*

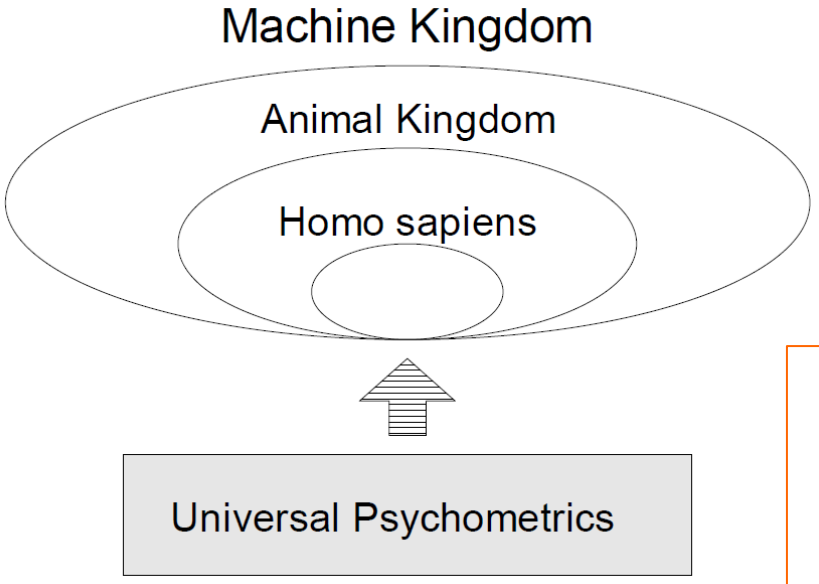
UNIVERSAL PSYCHOMETRICS

- Evaluation is always harder the less we know about the subject.
  - The less we take for granted about the subjects the more difficult it is to construct a test for them.
    - Human intelligence evaluation (psychometrics) works because it is highly specialised for humans.
    - Animal testing works (relatively well) because tests are designed in a very specific way to each species.

Who would try to tackle a more general problem (evaluating *any* system) instead of the actual problem (evaluating machines)?

UNIVERSAL PSYCHOMETRICS

- The *actual* problem is the *general* problem:
  - What about 'animats'? And hybrids? And collectives?



Machine kingdom: any kind of individual or collective, either artificial, biological or hybrid.

Universal Psychometrics (Hernández-Orallo et al, 2014, Cog Sci Res) is the analysis and development of measurement techniques and tools for the evaluation of cognitive abilities of subjects in the machine kingdom.

UNIVERSAL PSYCHOMETRICS

- Elements:
  - Subjects: physically computable (resource-bounded) interactive systems.
  - Cognitive task: physically computable interactive systems with a score function.
  - Cognitive ability (or task class): set of cognitive tasks.
    - The separation between task-specific and ability-specific becomes a progressive thing, depending on the generality of the class.
  - Interfaces: between subjects and tasks (observations-outputs, actions-inputs), score-to-reward mappings.
  - Distributions over a task class
    - performance as average case performance on a task class.
  - Difficulty functions computationally defined from the task itself.
- Some of these elements found in psychometrics and comparative cognition
  - Overhauled and founded here with the theory of computation and AIT.
- Tests can be universal or not, depending on the application.
- *Strong objections are understandable.*

CONCLUSIONS

- Two views of AI evaluation
  - Task-oriented evaluation
    - Still a huge margin of improvement in the way AI systems are evaluated.
    - The key issues are *M* and *p*, and distinguishing the definition of the problem class from an effective sampling procedure (testing procedure).
  - Ability-oriented evaluation
    - The notion and evaluation of ability is more elusive than the notion of task.
    - Scattered efforts in AI, psychometrics, AIT and comparative cognition:
      - Universal psychometrics as a unified view for evaluation of cognitive abilities.
- More a matter of degree as sets of tasks become wider.

CONCLUSIONS

- AI evaluation has not been a priority for AI in the past.
  - Not even recognised as an imperative problem or mainstream research.
- Measuring intelligence is a key ingredient for understanding what intelligence is (and, of course, to devise intelligent artefacts).
- Increasing need for system evaluation:
  - Plethora of bots, robots, artificial agents, avatars, control systems, 'animats', hybrids, collectives, etc., systems that develop and change with time.
  - Crucial for the *technological singularity* once (and if) achieved.
- A challenging problem…

> Artificial intelligence requires an accurate, effective, non-anthropocentric, meaningful and computational way of evaluating its progress, by evaluating its artefacts.

QUESTIONS?

Warning!
Intelligent answers not guaranteed.

* A paper version of this presentation, including full coverage of topics and references can be found at: http://arxiv.org/abs/1408.6908

- *Explorers* needed.
  - The machine kingdom is a space of cosmic dimension!

> "A smart machine will first consider which is more worth its while: to perform the given task or, instead, to figure some way out of it. Whichever is easier. And why indeed should it behave otherwise, being truly intelligent? For true intelligence demands choice, internal freedom. And therefore we have the malingerants, fudgerators, and drudge-dodgers, not to mention the special phenomenon of simulimbecility or mimicretinism. A mimicretin is a computer that plays stupid in order, once and for all, to be left in peace. And I found out what dissimulators are: they simply pretend that they're *not* pretending to be defective. Or perhaps it's the other way around. The whole thing is very complicated."
>
> Stanisław Lem, "The Futurological Congress (1971)"