Comparing Humans and AI Agents

Javier Insa-Cabrera[1], David L. Dowe[2], Sergio España-Cubillo[1], M.Victoria Hernández-Lloreda[3], José Hernández Orallo[1]

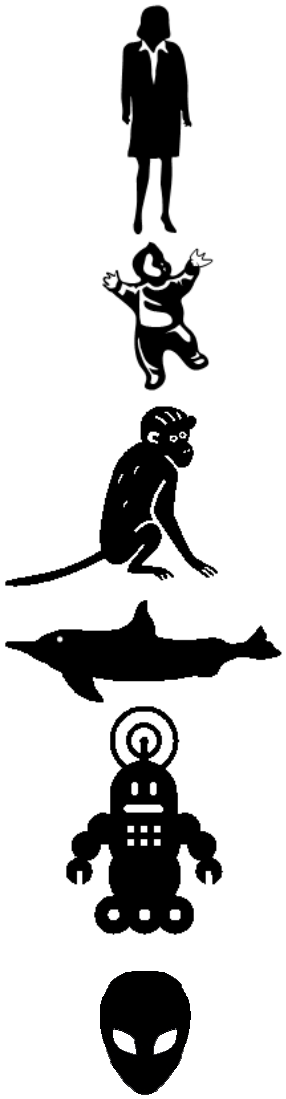
1. *Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, Spain.*
2. *Computer Science & Software Engineering, Clayton School of I.T., Monash University, Clayton, Victoria, 3800, Australia.*
3. *Departamento de Metodología de las Ciencias del Comportamiento, Universidad Complutense de Madrid, Spain*

AGI'2011 - Mountain View, California, 3-7 August 2011

Outline

- Measuring intelligence universally
- Precedents
- Test setting and administration
- Agents and interfaces
- Results
- Discussion

Measuring intelligence universally



- Can we construct a ‘universal’ intelligence test?

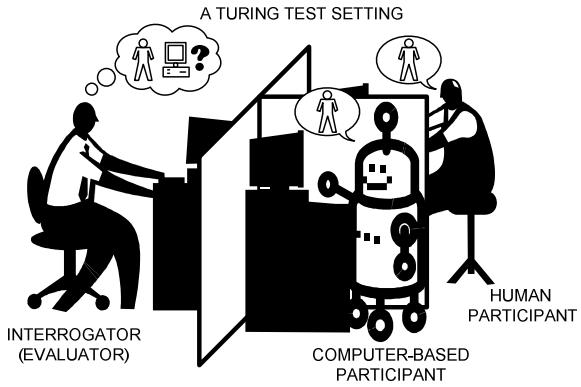
Project: **anYnt** (Anytime Universal Intelligence)

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

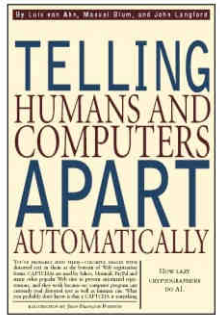
- Any kind of system (biological, non-biological, human)
- Any system now or in the future.
- Any moment in its development (child, adult).
- Any degree of intelligence.
- Any speed.
- Evaluation can be stopped at any time.

Precedents

- Imitation Game "Turing Test" (Turing 1950):
  - It is a test of *humanity*, and needs human intervention.
  - Not actually conceived to be a practical test for measuring intelligence up to and beyond human intelligence.



- CAPTCHAs (von Ahn, Blum and Langford 2002):
  - Quick and practical, but strongly biased.
  - They evaluate *specific* tasks.
  - They are not conceived to evaluate intelligence, but to tell humans and machines apart at the current state of AI technology.
  - It is widely recognised that CAPTCHAs will not work in the future (they soon become obsolete).



Precedents

- Tests based on Kolmogorov Complexity (compression-extended Turing Tests, Dowe 1997a-b, 1998) (C-test, Hernandez-Orallo 1998).
  - Look like IQ tests, but formal and well-grounded.
  - Exercises (series) are not arbitrarily chosen.
  - They are drawn and constructed from a universal distribution, by setting several 'levels' for *k*:

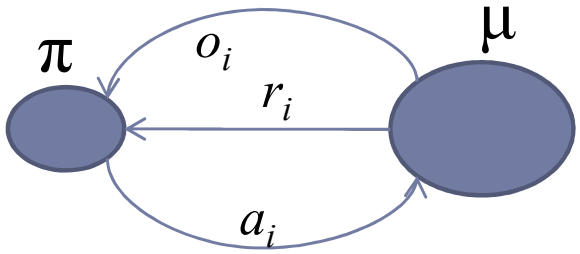
$k = 9$: a, d, g, j, . . . Answer : m
$k = 12$: a, a, z, c, y, e, x, . . . Answer : g
$k = 14$: c, a, b, d, b, c, c, e, c, d, . . . Answer : d

- However...
  - Some relatively simple algorithms perform well in IQ-like tests (Sanghi and Dowe 2003).
  - They are static (no planning abilities are required).

Precedents

- Universal Intelligence (Legg and Hutter 2007): an *interactive* extension to C-tests from sequences to environments.

$$\Upsilon(\pi, U) := \sum_{\mu=i}^{\infty} p_U(\mu) \cdot V_\mu^\pi = \sum_{\mu=i}^{\infty} p_U(\mu) \cdot E\left(\sum_{i=1}^{\infty} r_i^{\mu,\pi}\right)$$



= performance over a universal distribution of environments.

- Universal intelligence provides a definition which adds interaction and the notion of "planning" to the formula (so intelligence = learning + planning).
  - This makes this apparently different from an IQ (static) test.

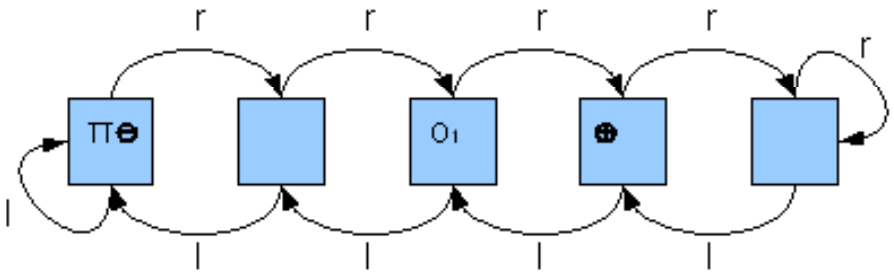
Precedents

- A definition of intelligence does not ensure an intelligence test.
- Anytime Intelligence Test (Hernandez-Orallo and Dowe 2010):
  - An interactive setting following (Legg and Hutter 2007) which addresses:
    - Issues about the difficulty of environments.
    - The definition of discriminative environments.
    - Finite samples and (practical) finite interactions.
    - Time (speed) of agents and environments.
    - Reward aggregation, convergence issues.
    - Anytime and adaptive application.
- An environment class Λ (Hernandez-Orallo 2010) (AGI-2010).

In this work we perform an implementation of the test and we evaluate humans and a reinforcement learning algorithm with it, as a proof of concept.

Test setting and administration

- Implementation of the environment class :
  - Spaces are defined as fully connected graphs.
  - Actions are the arrows in the graphs.
  - Observations are the ‘contents’ of each edge/cell in the graph.



  - Agents can perform actions inside the space.
  - Rewards:
- Two special agents *Good* (⊕) and *Evil* (⊖), which are responsible for the rewards. Symmetric behaviour, to ensure balancedness.

Test setting and administration

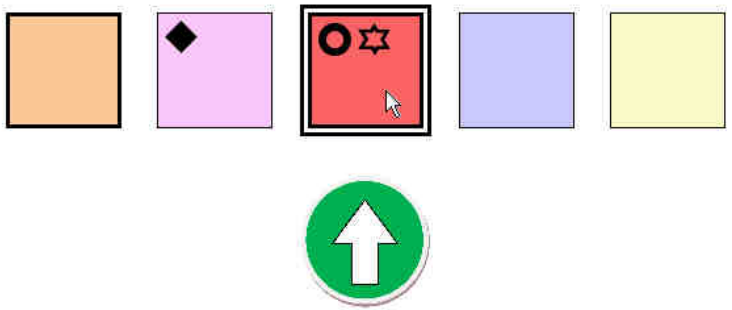
- We randomly generated only 7 environments for the test:
  - Different topologies and sizes for the patterns of the agents Good and Evil (which provide rewards).
  - Different lengths for each session (exercise) accordingly to the number of cells and the size of the patterns.

Env. #	No. cells (n_c)	No. steps (m)	p_{stop}
1	3	20	1/3
2	4	30	1/4
3	5	40	1/5
4	6	50	1/6
5	7	60	1/7
6	8	70	1/8
7	9	80	1/9
TOTAL	-	350	-

  - The goal was to allow for a feasible administration for humans in about 20-30 minutes.

Agents and interfaces

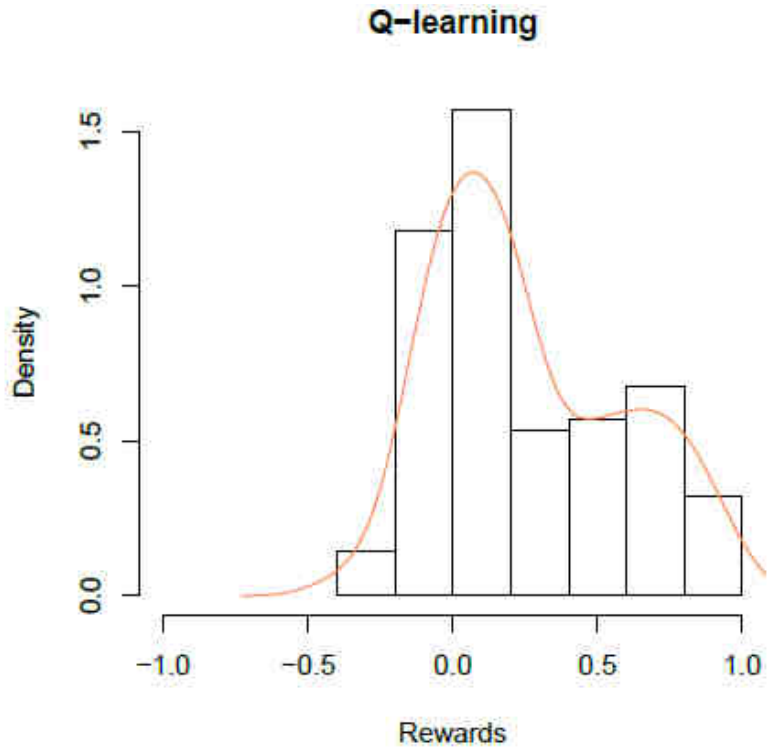
- An AI agent: Q-learning
  - A simple choice. A well-known algorithm.
- A biological agent: humans
  - 20 humans were used in the experiment
  - A specific interface was developed for them, while the rest of the setting was equal for both types of agents.

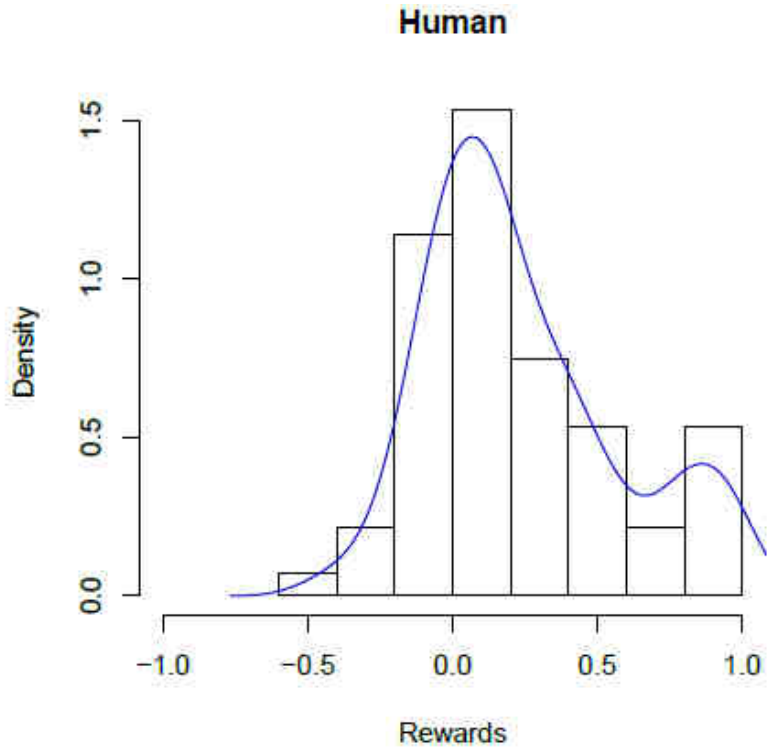


  - http://users.dsic.upv.es/proy/anynt/human1/test.html

Results

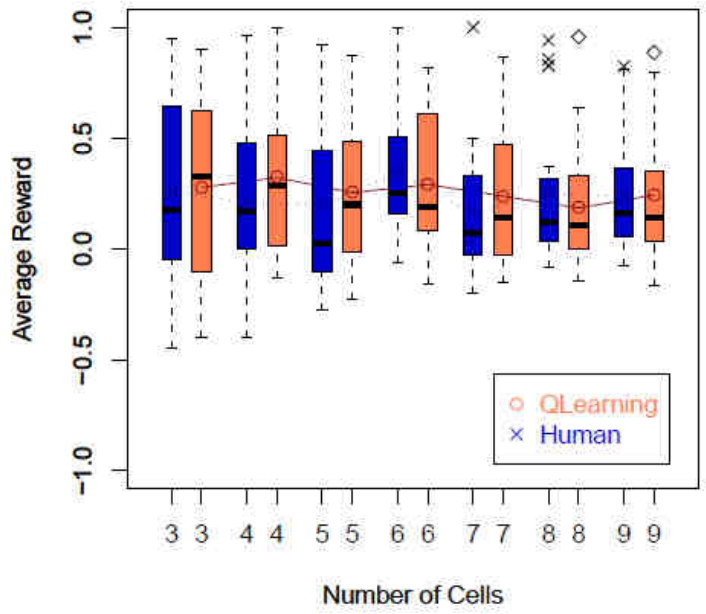
- Experiments were paired.
  - Results show that performance is fairly similar.

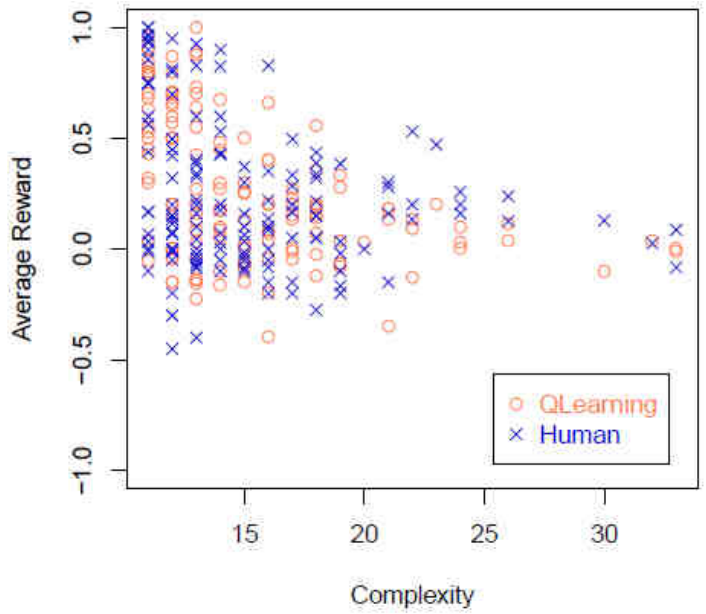




Results

- Analysis of the effect of complexity :
  - Complexity is approximated by using LZ (Lempel-Ziv) coding to the string which defines the environment.





  - Lower variance for exercises with higher complexity.
  - Slight inverse correlation with complexity (difficulty ↑, reward ↓).

Discussion

- Not many studies comparing human performance and machine performance on non-specific tasks.
  - The environment class here has not been designed to be anthropomorphic.
  - The AI agent (Q-learning) has not been designed to address this problem.

- The results are consistent with the C-test (Hernandez-Orallo 1998) and with the results in (Sanghi & Dowe 2003), where a simple algorithm is competitive in regular IQ tests.

Discussion

- The results show *this is not a universal intelligence test*.
  - The use of an interactive test has not changed the picture from the results in the C-test.
- What may be wrong?
  - A problem of the current implementation. Many simplifications made.
  - A problem of the environment class. Both this and the C-test used an inappropriate reference machine.
  - A problem of the environment distribution.
  - A problem with the interfaces, making the problem very difficult for humans.
  - A problem of the theory.
    - Intelligence cannot be measured universally.
    - Intelligence is factorial. Test must account for more factors.
    - Using algorithmic information theory to precisely define and evaluate intelligence may be insufficient.

Thank you!

Some pointers:

- Project: **anYnt** (Anytime Universal Intelligence)

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

- Have fun with the test

  http://users.dsic.upv.es/proy/anynt/human1/test.html