The ANYNT Project Intelligence Test Λ_{one}

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

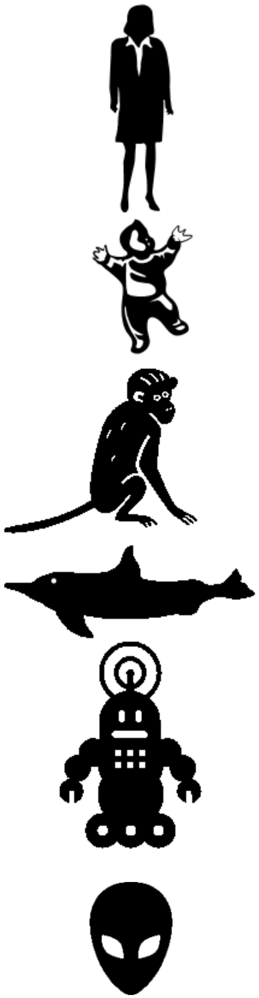
1. *Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, Spain.*
2. *Computer Science & Software Engineering, Clayton School of I.T., Monash University, Clayton, Victoria, 3800, Australia.*
3. *Departamento de Metodología de las Ciencias del Comportamiento, Universidad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

CQRW2012 - AISB/IA-CAP 2012 World Congress, July 4-5, Birmingham, UK

Outline

- Measuring intelligence universally
- Precedents
- Λ_{one} Test setting
- Testing AI performance
- Testing different systems
- Discussion

Measuring intelligence universally



- Can we construct a ‘universal’ intelligence test?

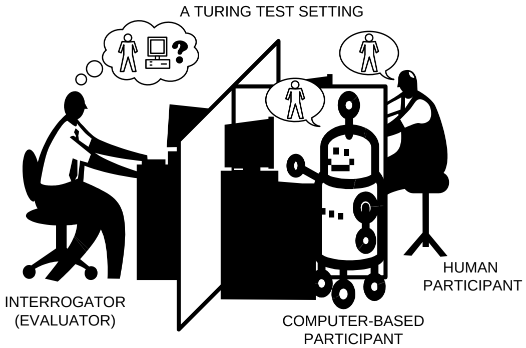
Project: **anYnt** (Anytime Universal Intelligence)

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

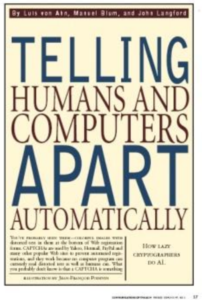
- Any kind of system (biological, non-biological, human).
- Any system now or in the future.
- Any moment in its development (child, adult).
- Any degree of intelligence.
- Any speed.
- Evaluation can be stopped at any time.

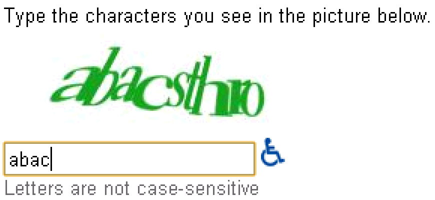
Precedents

- 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*:

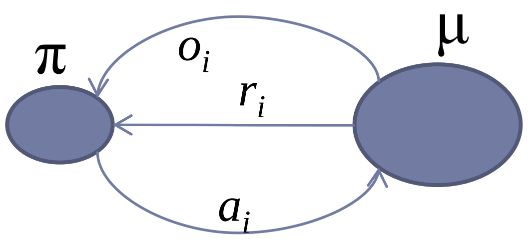
$$\begin{array}{lll} k=9 & : \text{a, d, g, j, } \ldots & \text{Answer : m} \\ k=12 & : \text{a, a, z, c, y, e, x, } \ldots & \text{Answer : g} \\ k=14 & : \text{c, a, b, d, b, c, c, e, c, d, } \ldots & \text{Answer : d} \end{array}$$

- 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.

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



= 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

- Kolmogorov Complexity

$$K_U(x) := \min_{p \ such \ that \ U(p)=x} l(p)$$

where l(p) denotes the length in bits of p and U(p) denotes the result of executing p on U.

- Universal Distribution

Given a prefixed-free machine U, the universal probability of string x is defined as:

$$p_U(x) := 2^{-K_U(x)}$$

Precedents

- ## Levin's Kt Complexity

$$Kt_U(x) := \min_{p \ such \ that \ U(p)=x} \{l(p) + \log time(U, p, x)\}$$

where l(p) denotes the length in bits of p and U(p) denotes the result of executing p on U, and time(U,p,x) denotes the time that U takes executing p to produce x.

- ## Time-weighted Universal Distribution

Given a prefix-free machine U, the universal probability of string x is defined as:

$$p_U(x) := 2^{-Kt_U(x)}$$

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).

Λ_{one} Test setting

- Discriminative environments.
- Interact infinitely: Must be a pattern (Good and Evil).

- Balanced environments.
  - Symmetric rewards.

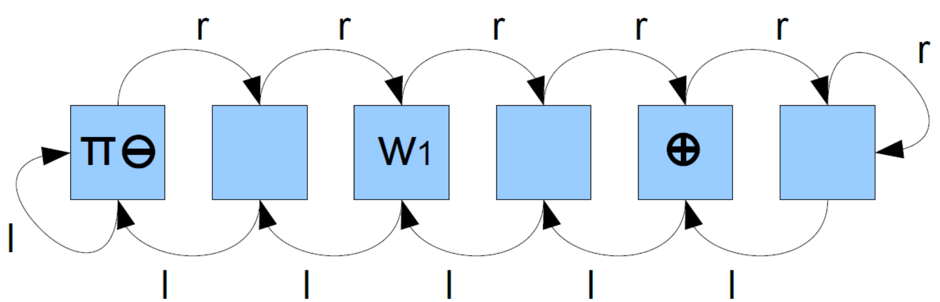
$$\forall i : -1 \leq r_i \leq 1$$

  - Symmetric behaviour for Good and Evil.

- Agents have influence on rewards: Sensitive to agents' actions.

Λ_{one} Test setting

- 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.

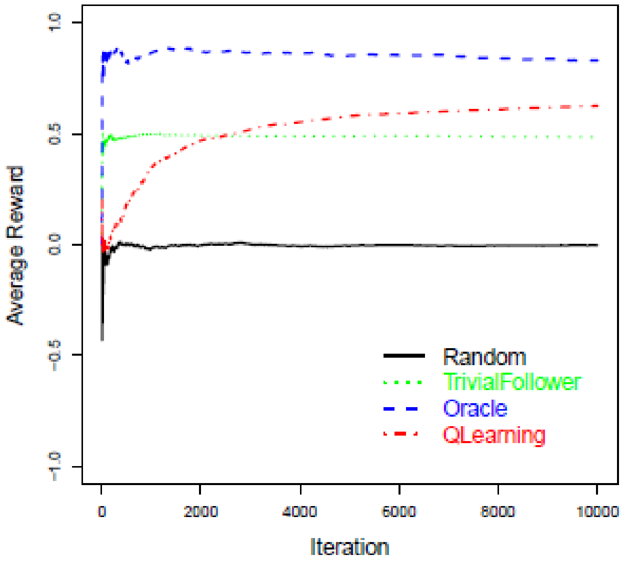


    - Agents can perform actions inside the space.
    - Rewards: Two special agents Good (⊕) and Evil (⊖), which are responsible for the rewards.

Testing AI performance

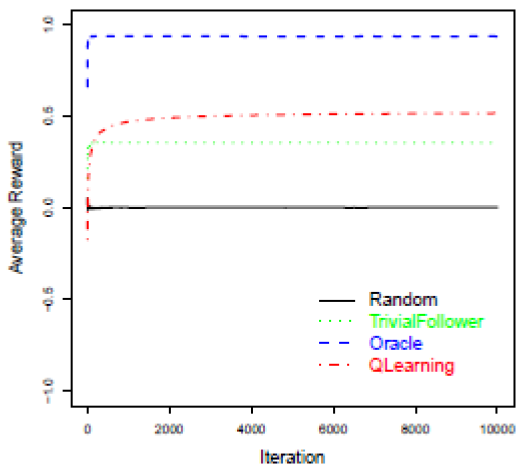
- Test with 3 different complexity levels (3,6,9 cells).
  - We randomly generated 100 environments for each complexity level with 10,000 interactions.
  - Size for the patterns of the agents Good and Evil (which provide rewards) set to 100 actions (on average).

- Evaluated Agents:
  - **Q-learning**
  - Random
  - Trivial Follower
  - Oracle

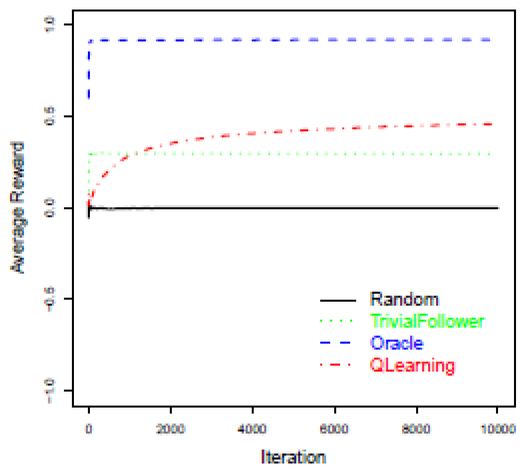


Testing AI performance

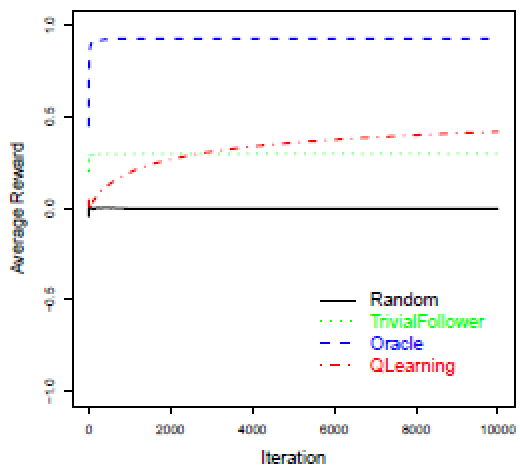
- Experiments with increasing complexity.
  - Results show that Q-learning learns slowly with increasing complexity.



3 Cells



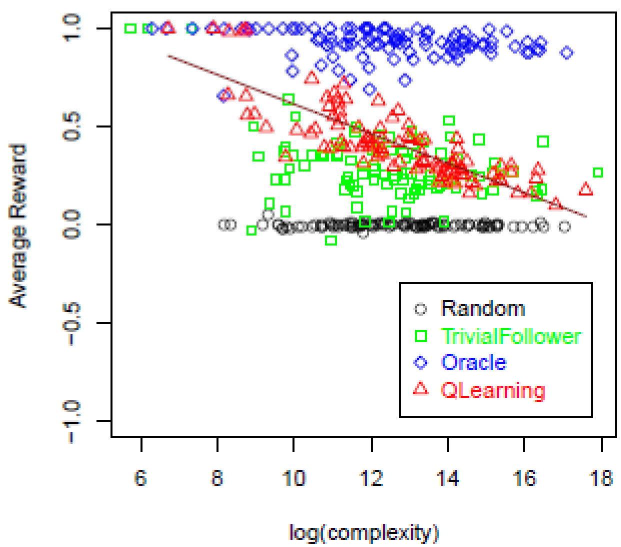
6 Cells



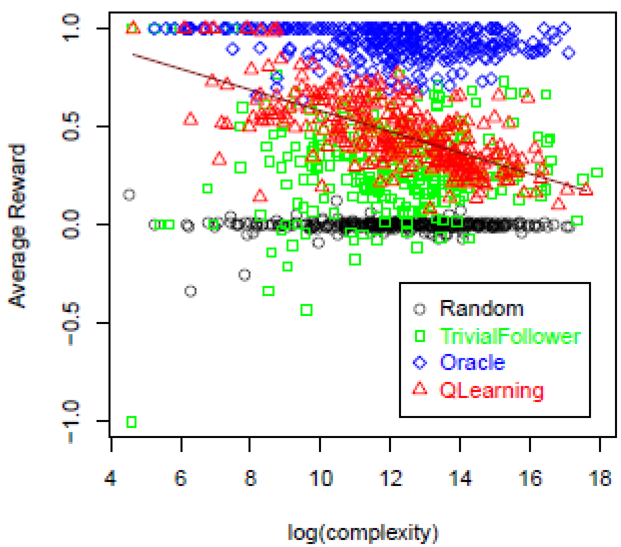
9 Cells

Testing AI performance

- Analysis of the effect of complexity:
  - Complexity of environments is approximated by using (Lempel-Ziv) LZ(concat(S,P)) x |P|.



9 Cells



All environments

  - Inverse correlation with complexity (difficulty ↑, reward ↓).

Testing different systems

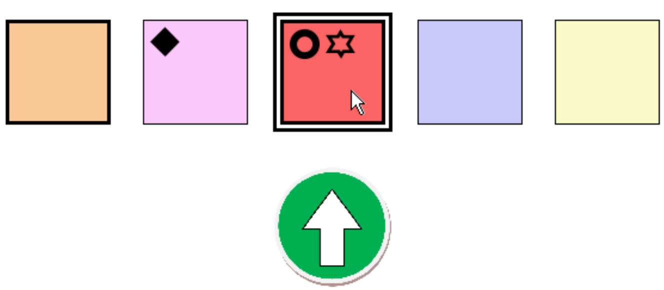
- Each agent must have an appropriate interface that fits its needs (Observations, actions and rewards):

- AI agent

b:E:πGa::

+1.0

- Biological agent: 20 humans



Testing different systems

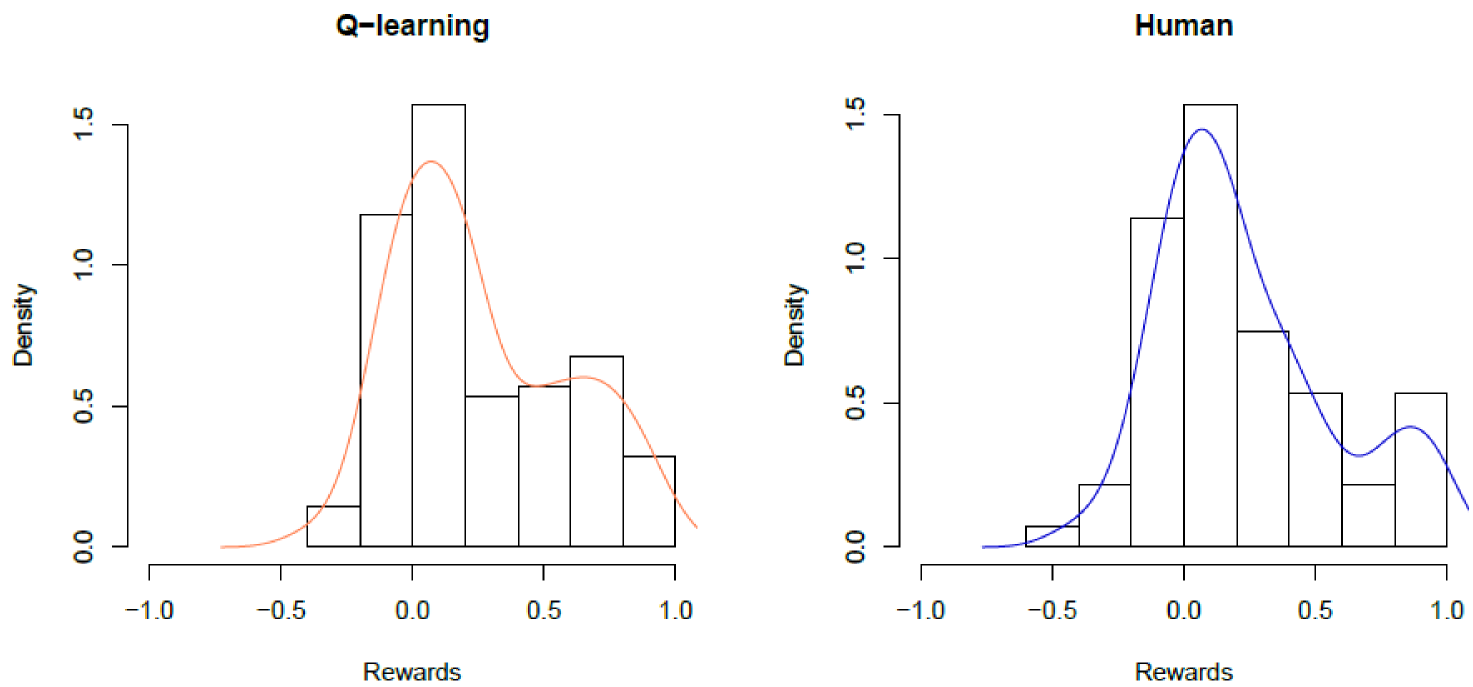
- We randomly generated only 7 environments for the test:
  - Different topologies and sizes for the patterns of the agents Good and Evil (which provide rewards).
  - Different lengths for each session (exercise) accordingly to the number of cells and the size of the patterns.

Env. #	No. cells (n_c)	No. steps (m)	Pattern length (on average)
1	3	20	3
2	4	30	4
3	5	40	5
4	6	50	6
5	7	60	7
6	8	70	8
7	9	80	9
TOTAL	-	350	-

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

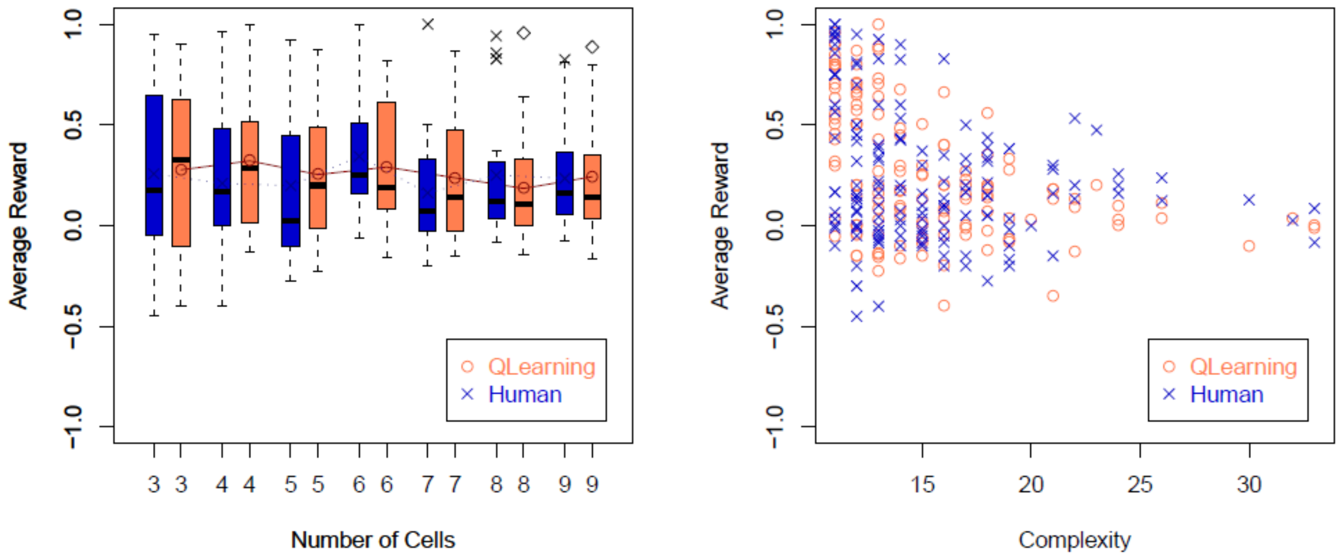
Testing different systems

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



Testing different systems

- 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

- Environment complexity is based on an approximation of Kolmogorov complexity and not on an arbitrary set of tasks or problems.
  - So it's not based on:
    - Aliasing
    - Markov property
    - Number of states
    - Dimension
    - …
- The test aims at using a Turing-complete environment generator but it could be restricted to specific problems by using proper environment classes.
- An implementation of the Anytime Intelligence Test using the environment class Λ can be used to evaluate AI systems.

Discussion

- The test is not able to evaluate different systems and put in the same scale. The results show *this is not a universal intelligence test*.
- What may be wrong?
  - A problem of the current implementation. Many simplifications made.
  - A problem of the environment class.
  - 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.

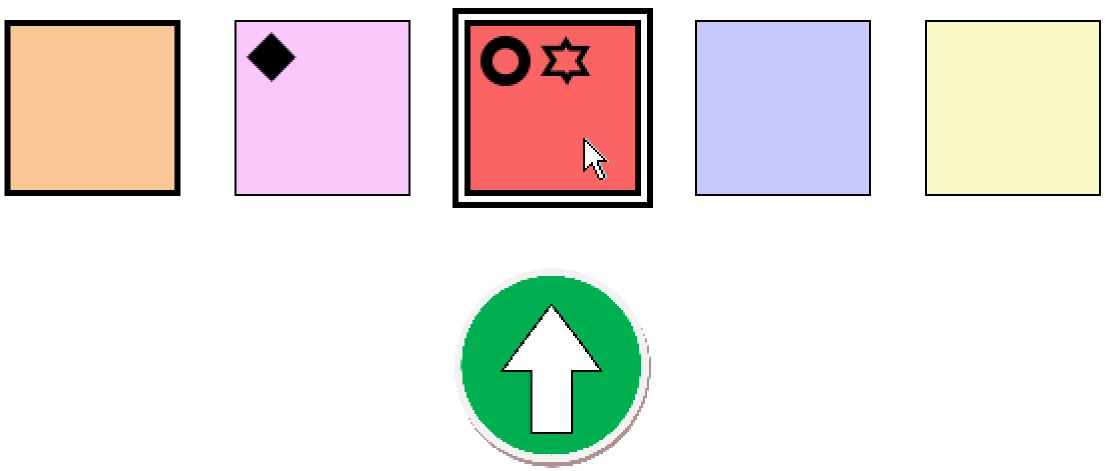
Thank you!

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