On Evaluating Agent Performance in a Fixed Period of Time

José Hernández Orallo

ELP-DSIC,
Universitat Politècnica de València,
Spain
jorallo@dsic.upv.es

AGI-2010, Lugano, March 6th, 2010.

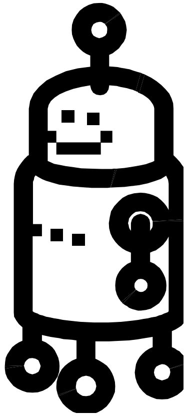
Not so dumb!

FROM: Herrmann, E., Call, J., Hernández-Lloreda, M.V., Hare, B., Tomasell, M. “Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis”, Science, 7 September 2007, Vol. 317. no. 5843, pp. 1360 - 1366, DOI: 10.1126/science.1146282.

Not so brilliant!

FROM: Herrmann, E., Call, J., Hernández-Lloreda, M.V., Hare, B., Tomasell, M. “Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis”, Science, 7 September 2007, Vol. 317. no. 5843, pp. 1360 - 1366, DOI: 10.1126/science.1146282.

Evaluating intelligence. Some issues.



1. Harder the less we know about the examinee.

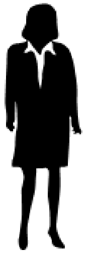
2. Harder if the examinee does not know it is a test.

3. Harder if evaluation is not interactive (static vs. dynamic).

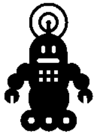
4. Harder if examiner is not adaptive.

Different subjects, different tests

- IQ tests:
  1. Human-specific tests. Natural language assumed.
  2. The examinees know it is a test.
  3. Generally non-interactive.
  4. Generally non-adaptive (pre-designed set of exercises)
- Other tests exist (interviews, C.A.T.)



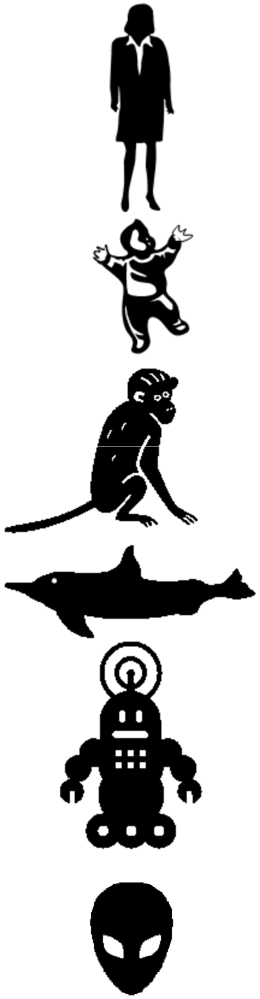
- Turing test:
  1. Held in a human natural language.
  2. The examinees ‘know’ it is a test.
  3. Interactive.
  4. Adaptive.
- Other task-specific tests exist.
  - Robotics, games, machine learning.



- 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 (pre-designed set of exercises).

- Animal 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 (pre-designed set of exercises).

Can we construct a test for all of them all?



- Without knowledge about the examinee,
- No natural language needed,
- Non-biased and without human intervention,
- Meaningful,
- Practical, and
- **Anytime.**

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

- Any system, now (human, non-human) 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

- Turing Test (Turing 1950): anytime and adaptive, but it is a test of humanity, and needs human intervention.
- Tests based on Kolmogorov Complexity (compression-extended Turing Tests, Dowe and Hajek 1998) (C-test, Hernandez-Orallo 1998). Very much like IQ tests, but formal and well-grounded. However, they can be cheated (Sanghi and Dowe 2003) and they are static.
- Captchas (von Ahn, Blum and Langford 2002): quick and practical, but strongly biased. They soon become obsolete.
- Universal Intelligence (Legg and Hutter 2007): can be seen as an interactive extension to C-tests, not as a test definition but as a theory of intelligence. However, a practical instance is hard to implement (computability problems, environment classes, time, ...).

The previous approaches ignore time or just set a time limit for the whole set of exercises.

Considering Time

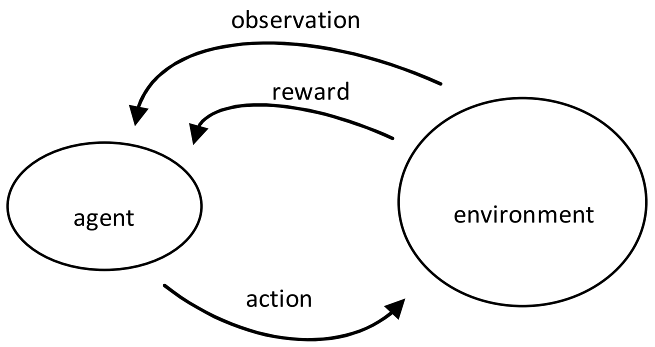
- Time is a key issue in measuring intelligence (and in performance in general).
  - A key issue in creating AGI systems. If time is not considered, the task is clearly much easier than it really is.
  - We can only ignore it when the agents behave in a similar time scale: adult humans, children, non-human animals, …
  - The use of a “virtual” or discrete time is the solution in some cases, but it is not when we do not know the time scale of the system to be evaluated.

How can we evaluate diverse subjects (fast and slow) with the same setting in a reasonable time?

This is a piece in the puzzle for an anytime test...

Evaluation Setting

- Interactive evaluation of an agent which does not know the goal of a test is based on the classical setting:



- Used in child evaluation, animal evaluation and, of course, in reinforcement learning.
- Reaction times are not generally considered.
- Reinforcement learning: discrete time frequently assumed.

Change of setting: discrete time on the environment, continuous time on the agent.

Classical Payoff Functions

- Total rewards:

$$V_\mu^\pi \Uparrow \tau := E\left(\sum_{i=1}^{n_\tau} r_i\right)$$

- Average rewards:

$$v_\mu^\pi || \tau := E\left(\frac{1}{n_\tau}\sum_{i=1}^{n_\tau} r_i\right)$$

- Discounted rewards (following Hutter 2006):

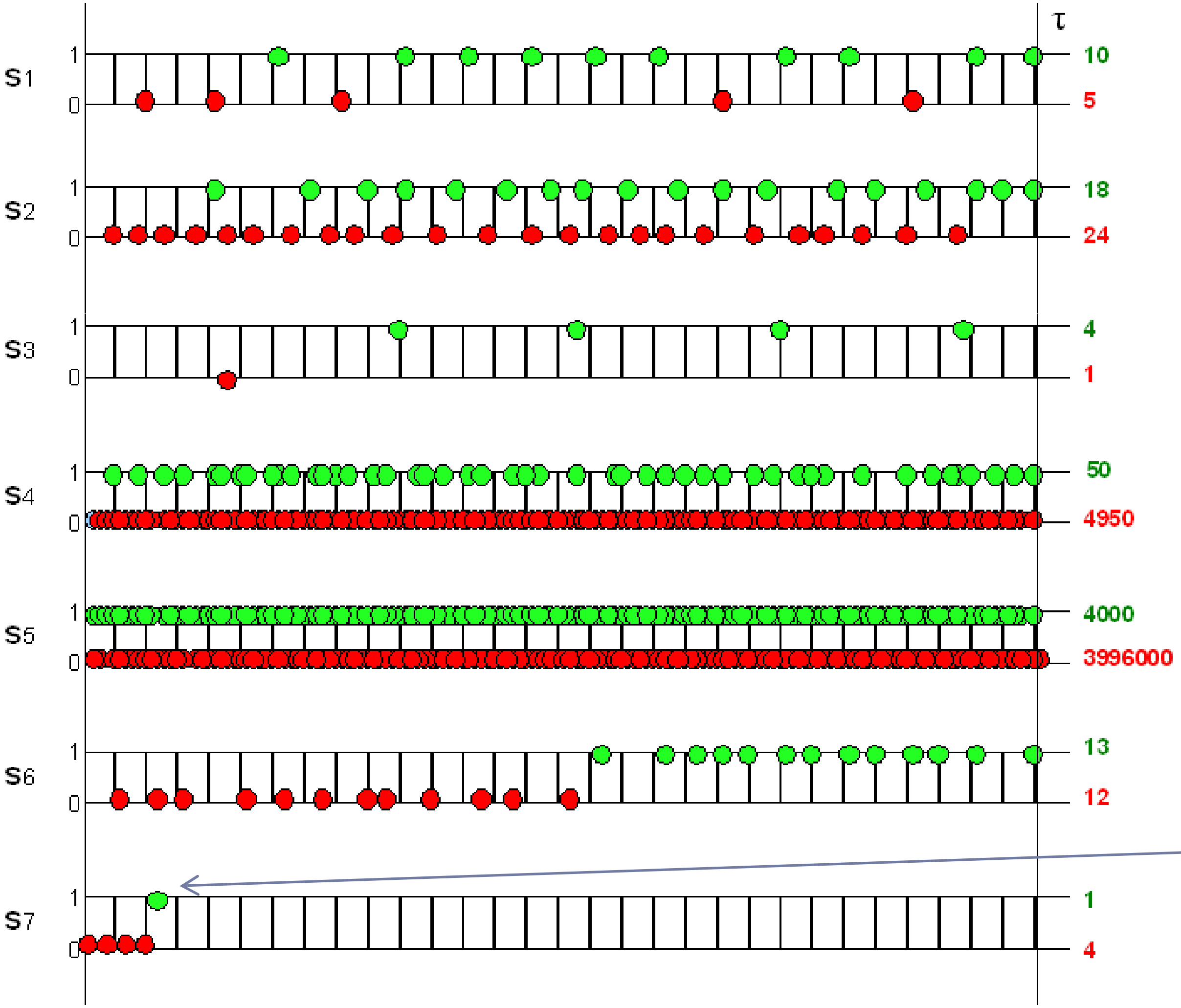
$$V_\mu^\pi |\gamma| \tau := E\left(\frac{1}{\Gamma^{n_\tau}}\sum_{i=1}^{n_\tau} \gamma_i r_i\right)$$

  - To avoid the arbitrary choice of γ , (Legg and Hutter 2007) propose:
    - Reward-Bounded (-summable) environment:

$$lim_{\tau\to\infty} V_\mu^\pi \Uparrow \tau = \sum_{i=1}^{\infty} r_i \leq 1$$

Example

- Testing in a fixed period of time τ:



Several options for the payoff:

- Total reward: S5
- Average reward: S3
- Discounted reward: S4 or S5.
- Considering prompt stabilisation: S7.
- Considering a statistically significant stabilisation: S6.

This stop can be intentional (or not)

Problems

- Environments with imbalanced rewards (paradises or hells) favour hyperactive or passive behaviours.
  - If we only give positive rewards to a the chimpanzee or a child, then they try to act rashly to get more and more rewards.
  - Using discounted rewards or reward-bounded (-summable) environments does not solve the problem (from an evaluation point of view).
- Opportunistic use of time.
  - As in gambling/bandit problems, a random agent can modulate time:
    - Acting very quickly when the average reward so far is bad.
    - Stopping when the average reward is good.
  - The expected value when tossing a coin like this (optimal stopping) is 0.79 (not 0.5). This happens with both virtual and real time.
  - Any average reward value can be obtained with infinite speed.

Requirements

- Measuring performance in a time τ under the following setting:
  - The overall allotted evaluation time τ is variable and independent of the environment and agent.
  - Agents can take a variable time to make an action, which can also be part of their policy.
  - The environment must react immediately (no delay time computed on its side).
  - The larger the time τ the better the assessment should be (in terms of reliability). This would allow the evaluation to be anytime.
  - A constant rate random agent π^{r}_{rand} should have the same expected valued for every τ and rate *r*.
  - The evaluation must be fair, avoiding opportunistic agents, which start with low performance to show an impressive improvement later on (faked improvement), or that stop acting when they get good results (by chance or not).

Approach

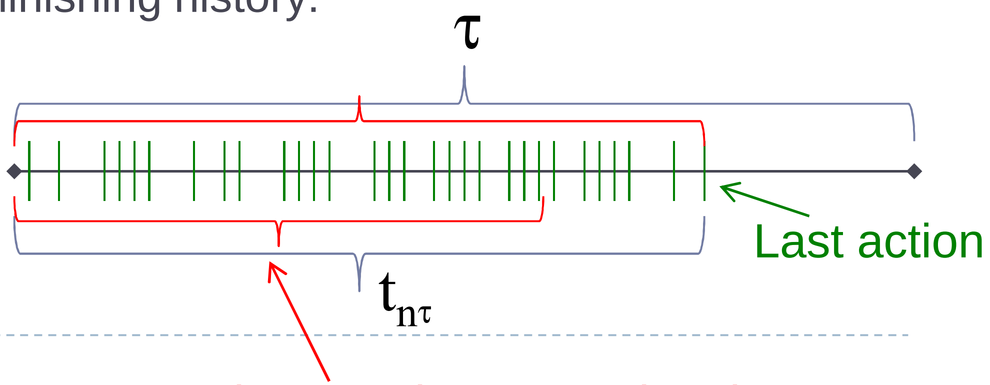
- Balanced environments: rewards go from –1 to 1 and:

$$\forall \tau > 0 \; : \; E\left(V_\mu^{\pi_{rand}^r} \Uparrow \tau\right) = E\left(\sum_{i=1}^{\lfloor r\times\tau\rfloor} r_i\right) = 0$$

  - The use of negative rewards is typical in economics, gambling and many games (everything that has been earned can be lost afterwards).
- Speed is not considered *in the payoff function*.
  - A fast agent would perform better because it can explore the environment faster (a different view of the exploitation vs. exploration dilemma).
- Correcting the measure using the last idle time.
  - Average reward per cycle with diminishing history:

$$\breve{v}_\mu^\pi \| \tau := E\left(\frac{1}{n^*}\sum_{i=1}^{n^*} r_i\right)$$

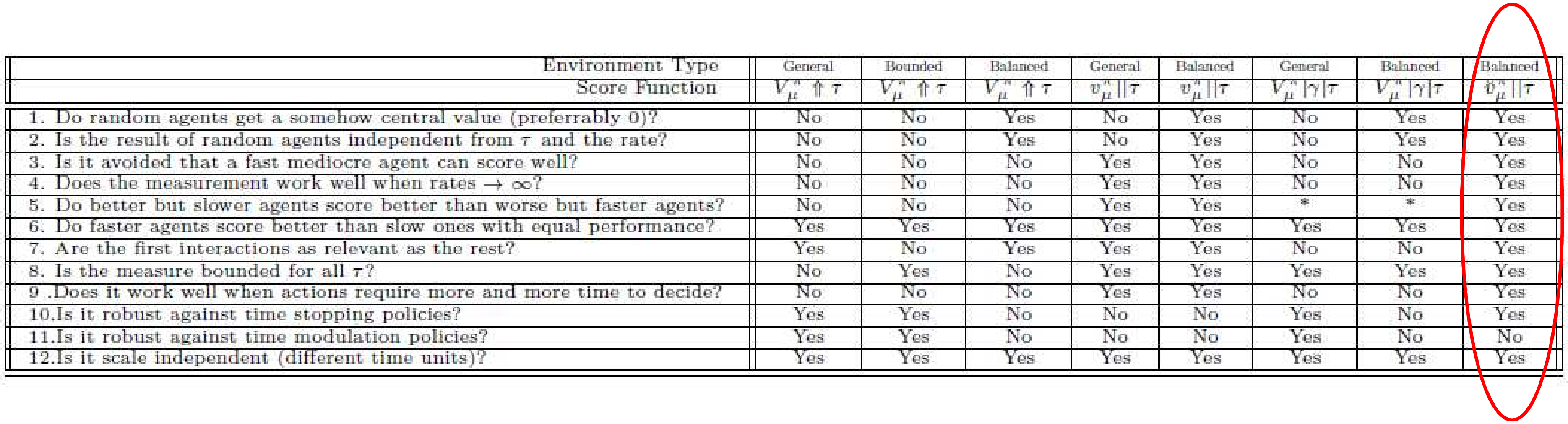
$$where \;\; n^* = \left\lfloor n_\tau \left(\frac{t_{n_\tau}}{\tau}\right)\right\rfloor$$



Only these actions are taken into account

Some Results

- With the adjusted payoff function, no stopping policy can make the expectation of a random agent better than 0 in a balanced environment.
  - This still allows for an intelligent use of time.
- Summary of payoff functions:



Environment Type	General	Bounded	Balanced	General	Balanced	General	Balanced	Balanced
Score Function	$V_\mu^\pi \Uparrow \tau$	$V_\mu^\pi \Uparrow \tau$	$V_\mu^\pi \Uparrow \tau$	$v_\mu^\pi \|\| \tau$	$v_\mu^\pi \|\| \tau$	$V_\mu^\pi \|\gamma\| \tau$	$V_\mu^\pi \|\gamma\| \tau$	$\check{v}_\mu^\pi \|\| \tau$
1. Do random agents get a somehow central value (preferrably 0)?	No	No	Yes	No	Yes	No	Yes	Yes
2. Is the result of random agents independent from τ and the rate?	No	No	Yes	No	Yes	No	Yes	Yes
3. Is it avoided that a fast mediocre agent can score well?	No	No	No	Yes	Yes	No	No	Yes
4. Does the measurement work well when rates $\rightarrow \infty$?	No	No	No	Yes	Yes	No	No	Yes
5. Do better but slower agents score better than worse but faster agents?	No	No	No	Yes	Yes	*	*	Yes
6. Do faster agents score better than slow ones with equal performance?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
7. Are the first interactions as relevant as the rest?	Yes	No	Yes	Yes	Yes	No	No	Yes
8. Is the measure bounded for all τ?	No	Yes	No	Yes	Yes	Yes	Yes	Yes
9 .Does it work well when actions require more and more time to decide?	No	No	No	Yes	Yes	No	No	Yes
10.Is it robust against time stopping policies?	Yes	Yes	No	No	No	Yes	No	Yes
11.Is it robust against time modulation policies?	Yes	Yes	No	No	No	Yes	No	No
12.Is it scale independent (different time units)?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Conclusions and Future Work

- We have addressed the performance evaluation in a finite period of time, considering that agent actions can take a variable time delay.
- The problem is apparently trivial, and it looks like the case when time is discrete and virtual (e.g. RL), but several problems appear.
  - Agents can become hyperactive if rewards are not balanced. Speed and not intelligence would be the key for a good result.
  - Agents can use an opportunistic use of time, by taking advantage of previous results (by chance or not), stopping and resting on their laurels.
- We propose the use of balanced environments and some simple modifications on the average reward which address the hyperactive and stopping problems.
  - This allows the setting to be used for anytime tests, where the more time is given, the higher the reliability of the measurement can be.
- Future work: continuous time for the environment to incorporate several agents of different speeds and capabilities *at the same time.*