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.
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.
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.

**Other task-specific tests exist.**
- Robotics, games, machine learning.

**Children’s intelligence evaluation:**
1. Perception and action abilities assumed.
2. The examinees do not know it is a test. Rewards are used.
3. Interactive.
4. Frequently 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?

- 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.
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.

Captcha (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.

- We adapt the classical setting:
  - We use a fixed time $\tau$, not a fixed number of interactions.
  - We consider reaction times.
  - Discrete time on the environment, continuous time on the agent.
Classical Payoff Functions

- Total rewards:
  \[ V^\pi_\mu \uparrow \tau := E \left( \sum_{i=1}^{n_\tau} r_i \right) \]

- Average rewards:
  \[ v^\pi_\mu || \tau := E \left( \frac{1}{n_\tau} \sum_{i=1}^{n_\tau} r_i \right) \]

- Discounted rewards (following Hutter 2006):
  \[ V^\pi_\mu |\gamma| \tau := E \left( \frac{1}{\Gamma n_\tau} \sum_{i=1}^{n_\tau} \gamma^i r_i \right) \]

- To avoid the arbitrary choice of \( \gamma \), (Legg and Hutter 2007) propose:
  - Reward-Bounded (-summable) environment:
    \[ \lim_{\tau \to \infty} V^\pi_\mu \uparrow \tau = \sum_{i=1}^{\infty} r_i \leq 1 \]
Several options for the payoff:

- Total reward: S5
- Average reward: S3
- Discounted reward: S4, S5 or S8.
- Considering prompt stabilisation: S7.
- Considering a statistically significant stabilisation: S6.

This stop can be intentional (or not)
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 (sticking) 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.
Balanced environments: rewards go from $-1$ to $1$ and:

$$\forall \tau > 0 : E \left( V_{\mu}^{\pi, rand} \uparrow \tau \right) = E \left( \sum_{i=1}^{\left\lfloor \tau \times \tau \right\rfloor} r_i \right) = 0$$

- 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:

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

where $n^* = \left\lfloor n_\tau \left( \frac{t_{n_\tau}}{\tau} \right) \right\rfloor$
With the adjusted payoff function, no stopping policy can make the expectation of a random agent different from 0 in a balanced environment.

This still allows for an intelligent use of time.

Summary of payoff functions:
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 make an opportunistic use of time, by taking advantage of previous results (by chance or not), stopping and resting on their laurels. This happens because there is not a fixed number of interactions.

We have proposed 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.