Benchmarking Benchmarks:
What Should Count as ‘Success’ Towards a Benchmark?

Vincent C. Müller

EUCogII & III – Anatolia College/ACT

www.eucognition– www.typos.de

ICRA 2011, Shanghai, May 13th

1.Origin of the Question

What is a good benchmark? = what do we test? What is ‘cheating’?

1.1. Overview

1. EUCogII work on ‘Challenges’ for artificial cognitive systems: systematically motivated benchmark challenges
2. Notes on ‘experiments’, ‘cognition’, ‘behaviourism’ and ‘who cares?’
3. Explanation and defence of the thesis on how to handle ‘cheating’, what is ‘success’

2. EUCogII

2.1. Basics

Artificial Cognitive Systems & Robotics: Fragmentation and lack of a clear agenda

-> Community for Clarity (CC)

- FP7 ICT Cognitive Systems, Interaction, Robotics & related
- Ca. 30 events with over 1000 participants during EUCogII

Step back from their day-to-day research, talk to people other than the usual experts in their sub-field and consider the bigger picture.

- 720 members now, ca. 4 applications per week (15% rejection)

2.2. Main Events

- Hamburg “Challenges for artificial cognitive systems”
- Zürich “Learning and development”
- Palma “Multi-sensory integration”
- Thessaloniki “Embodiment - fad or future?”
- Groningen “Autonomous activity in real-world environments” 10.-11. Oct. 2011
- Vienna, EUCogIII & CogSys 2012, 22-24. Feb. 2012

3.‘Challenges’ in EUCogII (Rapperswil, Jan 2011)

- Starting point:

“The set of challenges should:

- provide a long-term vision and fruitful orientation for present work

- be theory and strategy neutral (not fashion dependent, open to new approaches)

- not be domain specific, not be oriented towards toy problems or scenarios

- be systematic

- be measurable”

3.1. A) We need systematically motivated benchmark challenges

- The tension between the two aims is inherent in “a long-term vision and fruitful orientation for present work”.
- We cannot expect to formulate benchmarks once and for all. (This is not mathematics; Hilbert’s 23 Problems.)

3.2. B) Benchmarks involve a two-dimensional space: a) measurable success *and* b) measurable variation in complexity of the environment

a) speed, quantifiable output, comparison to other agents (natural or artificial) or ‘quality’ of output (+ use of resources)

b) enumerate relevant factors or use probabilistic measures.

•Ability to establish clear comparable metrics is inversely proportional to the degree of achievement.

•Real environments can only be specified to a degree, i.e. cannot be formal.

3.3. C) Benchmark challenges must test an entire autonomous system in an environment

•System performance in particular abilities is strongly dependent on overall features of the system, involving a host of different abilities – even if the benchmarks measures one.

3.4. D) Benchmark challenges must specify ‘cheating’

If benchmark challenges are set with respect to success in an environment, we:

a) ignore internal workings and

b) allow any working solution

Illustrating the perils of benchmarking – cheating?

- I promised a system that would do x, my **demonstration** shows my system do x. Done!
- **Asimo** can walk up stairs. Done!
- **Stanley** drove autonomously. Done!
- **ACE** found its way from TUM to Marienplatz (with no map). Done!
- **RoboCup@Home** went shopping in a supermarket. Done!

To sum up:

FET Flagship “Robot Companion”? Done!

4. Notes

4.1. A Note on Experiments

“Towards Replicable Experiments in Robotics Research”

- **But:** We do engineering, not (primarily) natural science
- Replicability and predictability rely on the “uniformity of nature”
- Replicability in robotics under identical circumstances is given, the issues are

  a) Replicability under **different circumstances**

  b) Replicability with **different robots** (under identical circumstances)

4.2. A Historical Note on ‘Cognition’

- “Higher level cognition”, cognition/volition/emotion + sensing/acting
- Cognitive Science
  - Cognitive Science: function (information processing)
  - Cognitive Neuroscience: biological substrate (part of it)
- Not GOFAI but AI
  - “Intelligent systems”
  - Successfully pursues goals – flexible & robust
    - Not like our laptops
    - but rather like a cockroach
- **Cognitive system: Intelligent, flexible, often biologically inspired**

4.3. A Note on Behaviourism

Standard critique: Benchmarks test behaviour, not mental ability. (E.g. Turing test) – and showing pain behaviour is not being in pain (Behaviorism is false).

-> Aims may include mechanisms, not just behaviour

4.4. A Note of Caution: Good Benchmarks, Who Cares?

Who is interested in benchmarks that prove scientific success?

- not the scientists, really
- not the funding agencies, really
- not the public, really

5. Thesis: “Anything Goes” – There is no Cheating!

5.1. An Example: Stacking Toy Blocks (courtesy of Holk Kruse)

5.2. The Standard Response: Rulebooks (e.g. RoboCup)

We want to test a specific ability, so we prescribe in detail what is permitted and what is not. (We might hide some of the task from you to avoid pre-fabricated solutions.)

But:

- You cannot predict everything
- Prediction rules out creativity
- Habit is a crucial ability – Aristotle said
    - Evolution has no ‘Cheating’ – Adaptability Is just One Way
        - Humans are adaptable, but not well adapted
        - Most organisms are adapted, but not very adaptable

5.3. The perils of generalization (inductive inference)

- I promised a system that would do x, my demonstration shows my system do x. Done!
- Asimo can walk up stairs. Done!
- Stanley drove autonomously. Done!
- ACE found its way from TUM to Marienplatz (with no map). Done!
- RoboCup@Home went shopping in a supermarket. Done!

To sum up:

FET Flagship “Robot Companion”? Done!

5.4. Telling a Telling Joke

An philosopher, a physicist and a mathematician are on a train in Scotland and see a black sheep:

Philosopher: "How odd. Scottish sheep are black."

Physicist: “Some Scottish sheep are black."

Mathematician “At least one side of one sheep is black.”

[Philosopher: “ … appears black to me now.”]

6. Conclusions

1. Define benchmarks in terms of aims
2. Define these aims in terms of ability for success in a given environment and under variations of that environment.
3. Allow any solution, but be very careful in drawing general conclusions for other environments