Summary Report

Vincent C. Müller
15.3.2011

EUCogII Workshop "Challenges for Cognitive Systems" Rapperswil, 28-30 January 2011

http://www.eucognition.org/index.php?page=eucogii-workshop-challenges

Workshop Organizers: Vincent C. Müller, Toni Gomila

- This summary represents the views of the author. It does not necessarily reflect those of the workshop participants. -

I - Introduction

The workshop was set up to further the efforts within EUCogII towards the formulation and dissemination of 'Challenges for artificial cognitive systems'.

Following a single e-mail announcement to EUCogII members, 37 participants were selected from about twice as many applicants - essentially on the basis of seniority and a good mix. (The participants are listed on the site.)

Prior to the meeting, the participants were asked to fill in an online questionnaire, asking them to comment on proposals and to propose urgent issues. The results are published on the workshop site.

All but one participant agreed that the following formulations from the organisers provided a good 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"

II Narrative

Friday

The organisers introduced the subject and we had a discussion on the overall aims and strategy. We differentiated 'systemic challenges' and 'benchmark challenges', where the latter involves a metrics (but decided not to look at possible intermediate levels). We assumed that a cognitive system successfully pursues goals; it is flexible (adaptive) and autonomous - challenges can thus be defined as degree of success towards this aim.

Systemic challenges ordered by prominence in questionnaire results (pared down to groups):
1. learning and development
2. perception (and integration)
3. categorization, representation and inference
4. action and goal selection
5. communication and cooperation with other agents
6. distributed cognition and culture
7. experience (conscious or other)
8. motor coordination

We had a presentation and discussion of the euCognition (2006-8) effort on 'roadmap' construction, which highlighted the difficulty of that enterprise.

It was decided not to start not with 'systemic' but with benchmark challenges' and to allow participants to set up and join working groups to formulate benchmarks. Six workgroups were set up:
1. Evolution & development
2. Action selection
3. Categorization, representation & inference
4. Cooperation & Communication
5. Embodiment
6. "The whole iguana"

Saturday

Each of the working groups met 9:30-13:00 and presented their results (slides) in the plenary session after lunch. Despite the original decision to go straight for benchmarks all groups actually started by specifying what would constitute a good benchmark through identifying the systemic challenges that it would test.

In the afternoon, we had two presentations (Andringa and Malsburg) that deepened the understanding of our undertakings from a computer science and neuroscience point of view.

Sunday

After a summarizing introductory session, the same working groups met again to refine their proposals in the light of the work done so far, i.e. they were asked to specify:
- specific benchmark challenge(s) with metrics (variation of environments or of system success)
- the requisite abilities/systemic challenges
- pre-benchmarks, technologies that constitute 'progress towards' (e.g. in the next 5 years)

Results were presented in a closing discussion – this is a summary of the proposed benchmark challenges and their metrics, per group:

1. Evolution & development

a) Learn an arbitrary task within the same learning environment and to the same level as a biological creature. The biological creature will be drawn from an increasingly complex set of biological creatures.
b) Co-operative Auto-pilot (assistant) for vehicle driving. Individuals living, dying, sharing, discarding and redeploying experiences of "crash events" over many generations – metrics: "continuously" improving through evolution.

2. Action selection

An artificial cognitive system enters an unknown, possibly crowded supermarket, with a list of items to be bought for a human. (Multiple levels of action selection, multiple goals.) - metrics: complexity of environment, speed & quality of success.

3. Categorization, representation & inference

Place objects found in a room in the correct boxes on the table. - metrics: number of objects, complexity of categories, distractors, difficulty of placing objects

4. Cooperation & Communication

Cloud of humans, robots and machines - with distributed sensing, cognition, actuation, and communication across humans and machines, who give a small part of their resources in order to help towards a task; distributed and dynamically adapting to change of available resources. - metrics: user satisfaction & participation, success at given tasks

5. Embodiment

a) Play tennis - metrics: performance, energy consumption, program complexity, price
b) Human robot companion - metrics: trustworthiness, energy consumption, program complexity, price
(Note: Evaluation at systems level, not algorithms. Identify missing hardware & software technologies.)

6. “The whole iguana”
Benchmarking the success of a full cognitive agent within an environment. Since both degrees of success and complexity of environment are relevant factors, this can be seen as a two dimensional space with “environmental complexity" on one axis and “agent coping ability" on the other. In this space, systems that are closer to the ‘complexity’ axis can face a more complex environment (they need control); systems that are close to the ‘coping’ axis are more coping, more autonomous (they need more options). Technical benchmarks will involve several systemic challenges.

The results of group 6 reflect some of the general discussions, and there was some convergence on the human-based challenges as well (shopping, object collection and sorting, tennis player, ...).

General Remarks

The entire workshop was carried out in a hard-working and collaborative atmosphere, with all participants active in their groups, busy to produce an outcome that could be presented to the plenary. It was also a very successful networking event, where many connections were formed or reinforced, information exchanged, groups forged, projects planned, etc.

Three Rapperswil participants (Asfour, Bonsignorio, Müller) were invited to a strongly related „Workshop on Benchmarking and Evaluation of Interactive Cognitive Systems“, organised by the Excellence Cluster Cognitive Interaction Technology (CITEC), Bielefeld University (http://www.cit-ec.de/BCogS_2011/), 7 & 8.02.2011.

Bonsignorio and Müller are now organising a joint EUCogII-Euron Workshop on „Challenges, Good Experimental Methodology & Benchmarking“ at the euRobotics Forum 2011 (6-8.4.2011) in Västerås (http://www.eucognition.org/index.php?page=eucogii-euron-workshop)

Further EUCogII input is planned for the „Workshop about the road Towards Replicable Experiments in Robotics Research“ at ICRA 2011 (9-13.5.2011).

While the workshop presentations were not published, some material is available on the EUCogII Wiki:
http://www.eucognition.org/eucogii-wiki/EUCogII_Workshop_%22Challenges_for_Cognitive_Systems%22,_Rapperswil

III - Systematic Outcomes

[Note: These are fairly personal conclusions by the author; they may or may not be shared by the other participants.]

A) We need systematically motivated benchmark challenges
The tension between the two aims for a) measurable benchmarks and b) theoretically identifiable problems of the discipline remained throughout the workshop - it is inherent in the formulation of challenges that should provide both “a long-term vision and fruitful orientation for present work”.

One can try to specify systemic challenges in terms of systemic abilities (like “learning” or “perception”), but one is then a) committed to keeping a specific cognitive architecture constant and b) tends towards isolated benchmark challenges of these abilities - which often do not scale. So, we should aim for benchmark challenges and see systematic challenge as constraints on these.

While we should try to make benchmark challenges relatively theory-neutral, they will always reflect current theoretical concerns, and so these should be made explicit. We cannot expect to formulate benchmarks once and for all. (This is not mathematics; we cannot create something strictly analogous Hilbert's 23 Problems, but perhaps something as powerful as the Turing Test).

B) Benchmarks involve measurable success or measurable variation in complexity of the environment

The metrics for benchmark tests must be generated through measuring degrees of a) 'success' of the agent or, b) complexity of the environment (i.e. flexibility of the system). We might keep one of the factors constant, and measure the other, or measure variation in both. For 'success', we might have measures such as speed, quantifiable output, comparison to other agents (natural or artificial) or 'quality' of output. In addition to this, we might consider the necessary use of resources, esp. of computation time and of energy. For 'complexity', we might try to enumerate relevant factors or use probabilistic measures.

For both 'success' and 'complexity' the ability to establish clear comparable metrics is inversely proportional to the degree of achievement. Also, success in a given environment involves a degree of chance, especially if the environment can vary greatly (e.g. if it involves other agents). Real environments can only be specified to a degree, i.e. they cannot be formal environments. Both 'success' and 'complexity' allow the specification of benchmark challenges that become increasingly more difficult, i.e. that can be tackled at different points of scientific/technological development. Benchmark challenges can take the format of 'targeted competitions', with the related networking and PR effects.

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. We now know, for example, that we can't do successful vision without 'higher' cognition, but neither can we do it without 'lower' action-sensation and morphology (e.g. particular eyes or a turning head). The actual benchmarks proposed always involved a host of different abilities. Since we cannot isolate abilities, we must look at the whole system in its environment, at least in the longer run.

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.

There will thus be solutions to benchmark challenges that are undesirable for theoretical reasons and thus considered 'cheating' (typically the reason that they will not be transferable to different or more complex environments, including much ‚pre-fabricated' material). These can be 'outlawed' by architectural requirements or by task variation - in the latter case benchmarks be non-reproducible. Note that benchmark challenges do not in themselves provide a modeling of a given natural cognitive system - if these are required, the restrictions on what counts as cheating must be fairly narrow.