BENCHMARKING MEDICAL HRI, ROBOT THERAPY FOR THE DISABLED

ANDREA BONARINI

ARTIFICIAL INTELLIGENCE AND ROBOTICS LAB
DIPARTMENTO DI ELETTRONICA, INFORMAZIONE E BIOINGEGNERIA
POLITECNICO DI MILANO

E-MAIL: ANDREA.BONARINI@POLIMI.IT
URL:HTTP://WWW.DEIB.POLIMI.IT/PEOPLE/BONARINI

WHAT IS BENCHMARKING?

Objective performance evaluation of a device under controlled and replicable conditions

- Objective: independent from human judgement
  - Measurable quantities
  - (Statistical data analysis)
    subject to applicability conditions and experiment design
- Controlled conditions: all relevant features are known and controlled
  - What is relevant?
  - How is it possible to control?
- Replicable conditions
  - Controlled conditions
  - Possibility to replicate the conditions

Conditions quite hard to obtain for HRI systems

DANGER

Benchmarking originates from physical experiments, for which it has to be possible to satisfy the conditions

Any benchmarking procedure for a device is prone to methodological issues: we may like to certify a performance, possibly representative of the final use of the device (e.g., how well a washing machine washes), not only a characteristic of the device (e.g., whether in a standard condition an electrical shock occurs).

In many benchmarks, specific tests are selected and producers operate to pass the test, which might not be representative of the average operating conditions (e.g., washing machines), or might be faced in "special configurations" (e.g., cars (VW))

This is not acceptable for medical applications

WHY BENCHMARKING?

Objective evaluation … of what?

- Goal achievement (Yes/No result)
- Performance: an absolute value (possibly a measure), or a comparison with a standard device (comparative measure)

Improvement/development of the device (developers/researchers)

- Identification of quality/weakness
- Products (robots) or papers ;-)

Certification (Institutions that assign quality marks)

- Guarantee of repeatability in time and instances

PROBLEMS WITH BENCHMARKING IN HRI

What could we benchmark?

- Low level performance (e.g., word recognition)
  - Recorded data sets -> bias towards data sets: recorded, standard data are different from real operating conditions
  - Selected people that interacts -> bias towards the specific subjects
  - Robotic systems -> RRI is different from HRI

- High level performance (e.g., user involvement, therapeutic success)
  - Difficult to have standard data sets -> bias
  - Difficult to define goals in a precise way
  - Difficult to evaluate goal achievement and, even more, performance

A CHANGE IN PERSPECTIVE

Benchmarking is a problem in HRI, even more in medical applications, but…

Who does really need a device passing a benchmark with all these problems?

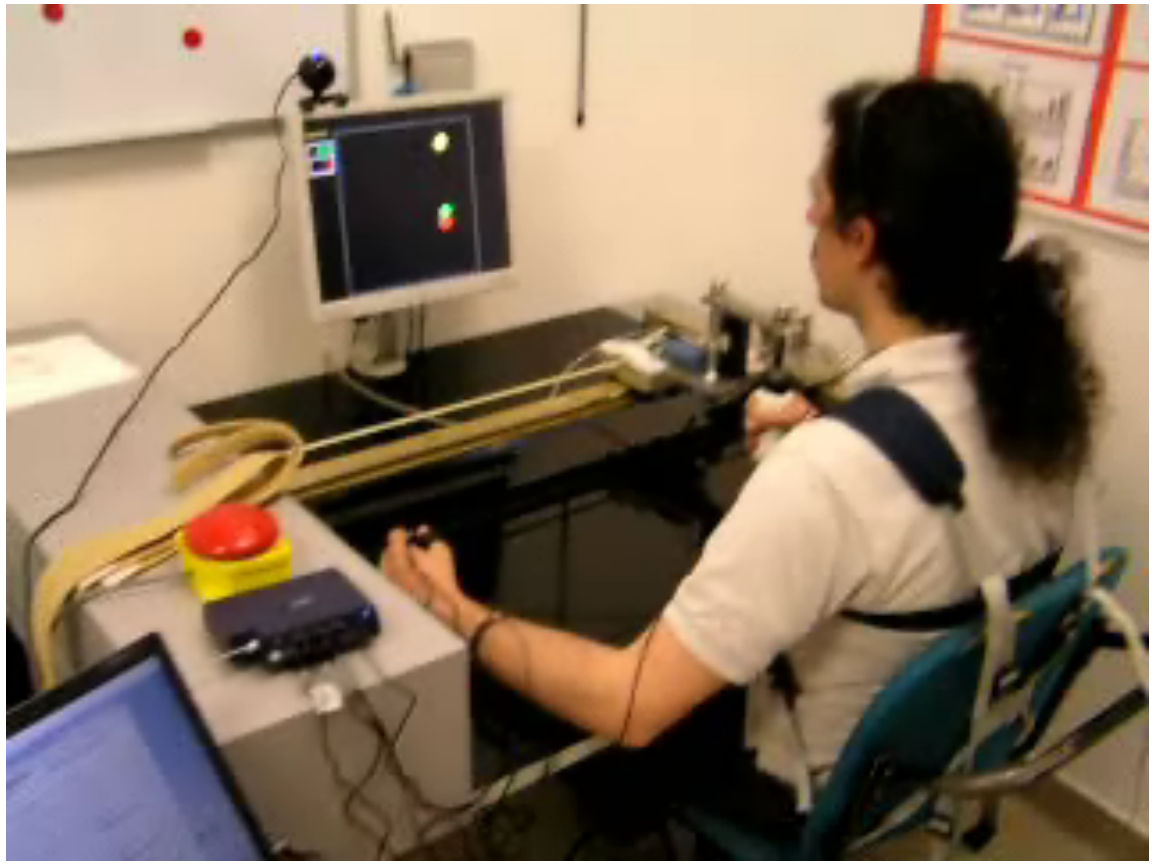
Wouldn't it be enough to guarantee basic characteristics, as done and accepted for most devices, and provide case-based evidence that the device is working fine?

Do not strive to replicate experiments, but measure goals achievements, in well-defined experimental conditions

Let's see a couple of case studies in the medical domain

ROBOT-BASED STROKE REHABILITATION

A patient with mobility limitation on the arm due to a stroke, exercises it by interacting with a robot that might help him/her when he/she cannot perform the prescribed movement



WHAT CAN WE BENCHMARK?

- Force that can be exerted by the robot
- Speed of the robot
- Control cycle (reaction) time

YES!!

- Acceptability of the interaction
- User's satisfaction
- Therapeutic success

NO benchmarking!!

Evaluation

- Lickert scales
- Users different from each other
- Description of the experimental setting
- Therapist's evaluation

Revision and improvement (affective detection and game setting)

ROBOT-BASED AUTISM REHABILITATION

Spatial and touch interaction with autistic children

When in autonomous mode, sensors detect children's interaction (hug, bump, punch, relative position,…) and the robot reacts by showing an emotion with sound, lights and movement

WHAT CAN WE BENCHMARK?

- Min/max speed of the robot
- Sensor data interpretation reliability
- Control cycle (reaction) time

- Acceptability of the interaction
- User involvement
- Therapeutic success

YES!!

NO benchmarking!!

Evaluation

- Users very different from each other
- Description of the (adaptive) experimental setting
- Therapist's evaluations (as for any other therapy)

Revision and improvement (stimuli adaptation, affective behavior design)

CONCLUSION

Benchmark only what can be properly benchmarked

Evaluate the rest, describing the case setting: **it's enough!**

Questions? Comments?

These ideas have been developed within the EU FP7 project RoCKIn