Credibility – An Enabling Factor for Trustworthiness

Univ.-Doz. DI Dr. Michael Hofbaur
JOANNEUM RESEARCH – ROBOTICS
ERF Workshop 21.03.2019

Trustworthy Robots?

trust /trust/, Noun:

1 Worthiness of being relied on
2 Confident expectation
3 A resting on the integrity,
4 …

- trust is basically the *willingness to be vulnerable* with the preconditions, that *risk* and *interdependence* exist
- so it's nothing that we can build into systems in engineering terms!

Safe Robots

In CE we trust!

- Is it enough to ensure CE conformity to gain trust in a potentially dangerous machine next to us?
- What about Cyber-Security issues that directly and physically interact with our world?
- Safety (& Security) is a prerequisite for trust!

Trustworthiness ←→ Credibility?

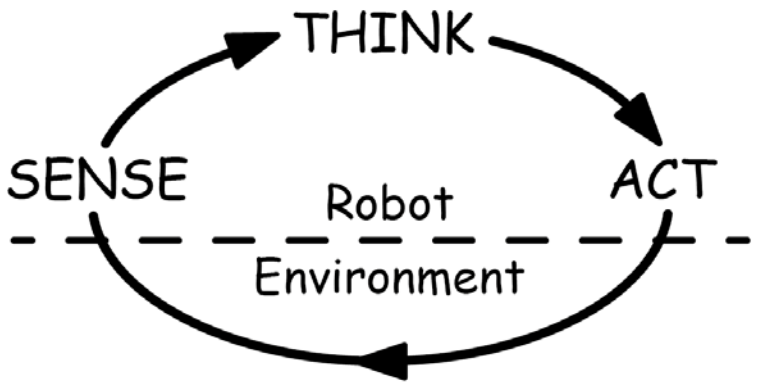
From the human's perspective:

- Obeying the rules is one prerequisite for gaining trust!
- Other aspects:
  - being predictable,
  - being accountable,
  - being capable of explaining / justifying one's actions
  - → being CREDIBLE qualifies to be a trustworthy partner!
- Trust is not a design property – to emphasize this: credibility!

How to design CREDIBLE robots that qualify a partners so that we can gain, maintain and never loose trust in them?

Credible Robots

- credibility from a robot‘s operational perspective:



- to my understanding, credibility as the basis for trust concerns all aspects of the control cycle, and beyond!

Credible SENSE

- Perceive all relevant aspects of the robot‘s environment

moreover

- credibly detect humans
- credibly anticipate their behaviour
- credibly anticipate the situational context
- credibly capture the safety context

- again, a sufficiently high integrity level is mandatory
- so it matters *what & how* we implement this functionality

Credible THINK

- credible supervisory control

- functional safety and cyber-security as a starting point
- plan a robot's actions that considers safety aspects
- derive predictable action sequence
- explainable automated reasoning & AI in general

- being accountable not just at the time of decision-making but also retrospectively

Credible ACT

- credible low-level control

- functional safety and cyber-security as a starting point
- safety-aware physical movements of the robot
- anticipatory behavioural patterns (movements, physical interactions, …)
- credible / anticipatory reflexive behaviour
- credible human-robot interaction (information)
- …

Summary

- trust is not an engineering property
- trust has to be gained and continuously maintained
- it‘s the humans that trust
- it‘s the robots that should be built in a credible way so that they qualify as being trustworthy

- *alongside with functional (safety) requirements, its human-factors that determine the framework for credibility*
- *credibility concerns all aspects of the control cycle, i.e. sense, think, act & remember*