UNIVERSITÀ DEGLI STUDI DI PALERMO

dipartimento di ingegneria unipa

CNR ISTC INSTITUTE OF COGNITIVE SCIENCES AND TECHNOLOGIES

ENDOWING ROBOTS WITH SELF-MODELING ABILITIES FOR TRUSTFUL HUMAN-ROBOT INTERACTIONS

Cristiano Castelfranchi, Antonio Chella, Rino Falcone, Francesco Lanza and Valeria Seidita

20TH WORKSHOP "FROM OBJECTS TO AGENTS"

OBJECTIVE

How to model and develop teammate robots performing trustful interaction with humans?

- Modeling and representing robot's knowledge
- The robot has decide and act in an autonomous fashion
- The robot has be self-adaptive

KEY IDEAS

- Triggering the decision process by means of attributing mental state to itself and to the others
- Integrating self-modeling and trust
- Employing BDI paradigm and Jason
  - extending BDI reasoning cycle for including self-modeling and justification

HUMAN-ROBOT TEAMING INTERACTION

- Two main situations:
  - Known and unchanging environment
  - Partially known and changing environment

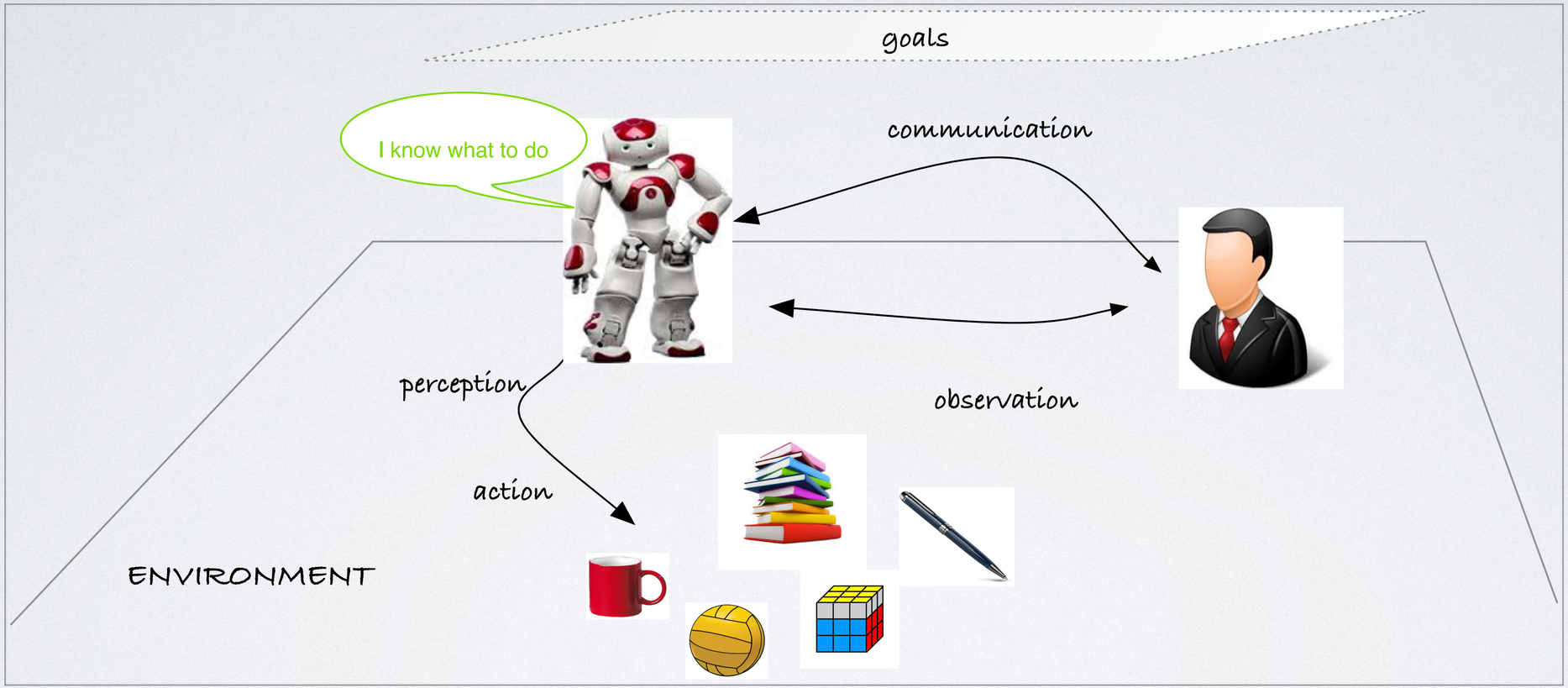
HUMAN-ROBOT TEAMING INTERACTION

- Known and unchanging environment
  - Each teammate knows everything before starting
  - Use knowledge for performing task and plans

HUMAN-ROBOT TEAMING INTERACTION

- Partially known and changing environment
  - Each teammate interacts mainly for:
    - Enhancing knowledge on the environment and on himself
    - Acquiring knowledge on what to do
  - Each teammate is:
    - aware of his own limitation and capabilities
    - establish a level of confidence in the other

HUMAN-ROBOT TEAMING INTERACTION

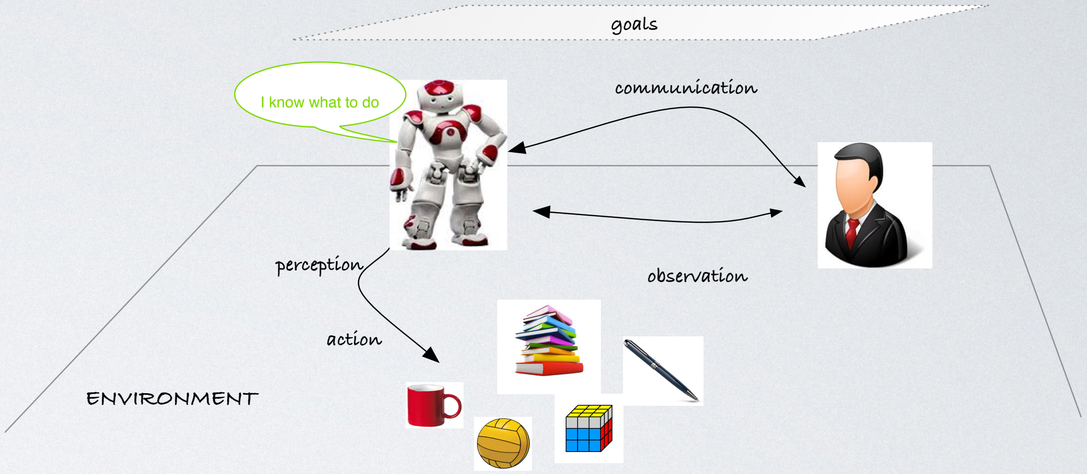


HUMAN-ROBOT TEAMING INTERACTION

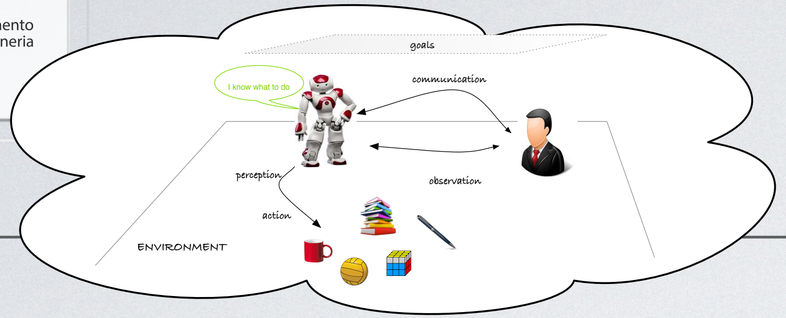
- Complex systems
  - where requirements are identified at runtime
    - changing environment conditions
    - presence of interacting users
  - global behavior emerges at runtime
- Need for exhibiting adaptation

A NEW PERSPECTIVE

- Changing perspective
- Robot is part of the environment



  - it senses itself as part of the environment along with all the other objects/agents



goals

I know what to do

communication

perception

observation

action

ENVIRONMENT

CHALLENGES

- Environment is not known a-priori
  - plans, tasks and all that is necessary for acting and deciding cannot be established at design time
- Equipping the robot with the ability **(at runtime)** to select the best action to perform
  - Knowledge acquisition
  - Decision-making process
- Several ingredients trigger the decision process:
  - goals, capabilities, mental states, emotions, **trust**…in general: awareness

SELF-MODELING AND TRUST

- The role of self-modeling and trust -> triggering the decision process
- The robot si able to create a model of the self
  - It is able to select an action on the base of what it knows about itself
  - it is able to decide which action to adopt

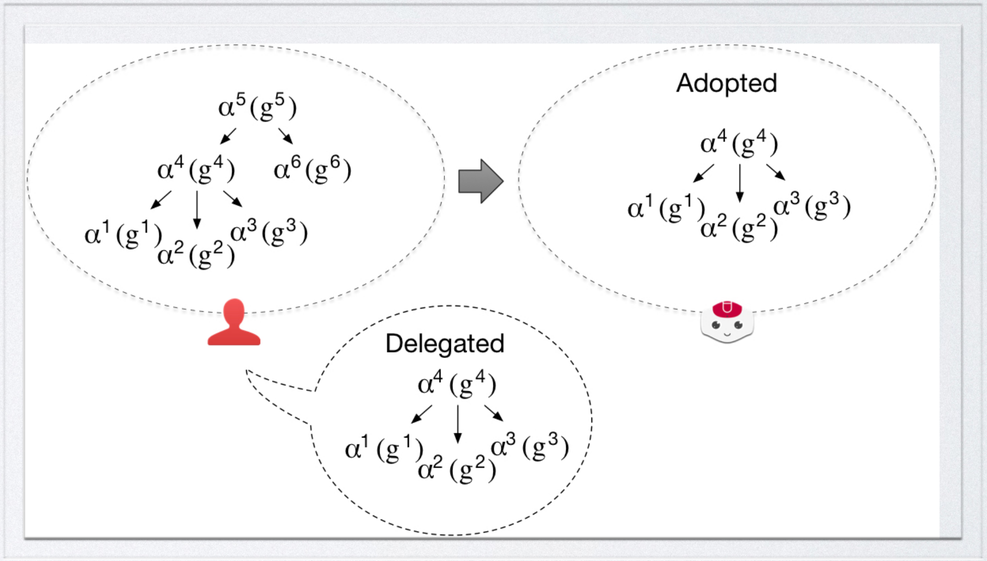
IMPLEMENTING SELF-MODELING ABILITIES

- Exploiting BDI practical reasoning and the trust model
- We extend the deliberation process

THE TRUST MODEL

- *The trust of a trustor agent in a trustee agent for a specific context to perform acts to realize the outcome result.*

$$TRUST(\mathbf{X}\ \mathbf{Y}\ \mathbf{C}\ \tau\ g_X)$$



THE BDI REASONING CYCLE

Robot has:
to commit or delegate actions and
reason on its own action

```
1.  B ← B0;          /* B0 are initial beliefs */
2.  I ← I0;          /* I0 are initial intentions */
3.  while true do
4.    get next percept ρ via sensors;
5.    B ← brf(B, ρ);
6.    D ← options(B, I);
7.    I ← filter(B, D, I);
8.    π ← plan(B, I, Ac); /* Ac is the set of actions */
9.    while not (empty(π) or succeeded(I, B) or impossible(I, B)) do
10.     α ← first element of π;
11.     execute(α);
12.     π ← tail of π;
13.     observe environment to get next percept ρ;
14.     B ← brf(B, ρ);
15.     if reconsider(I, B) then
16.       D ← options(B, I);
17.       I ← filter(B, D, I);
18.     end-if
19.     if not sound(π, I, B) then
20.       π ← plan(B, I, Ac)
21.     end-if
22.   end-while
23. end-while
```

MAPPING TRUST TO BDI

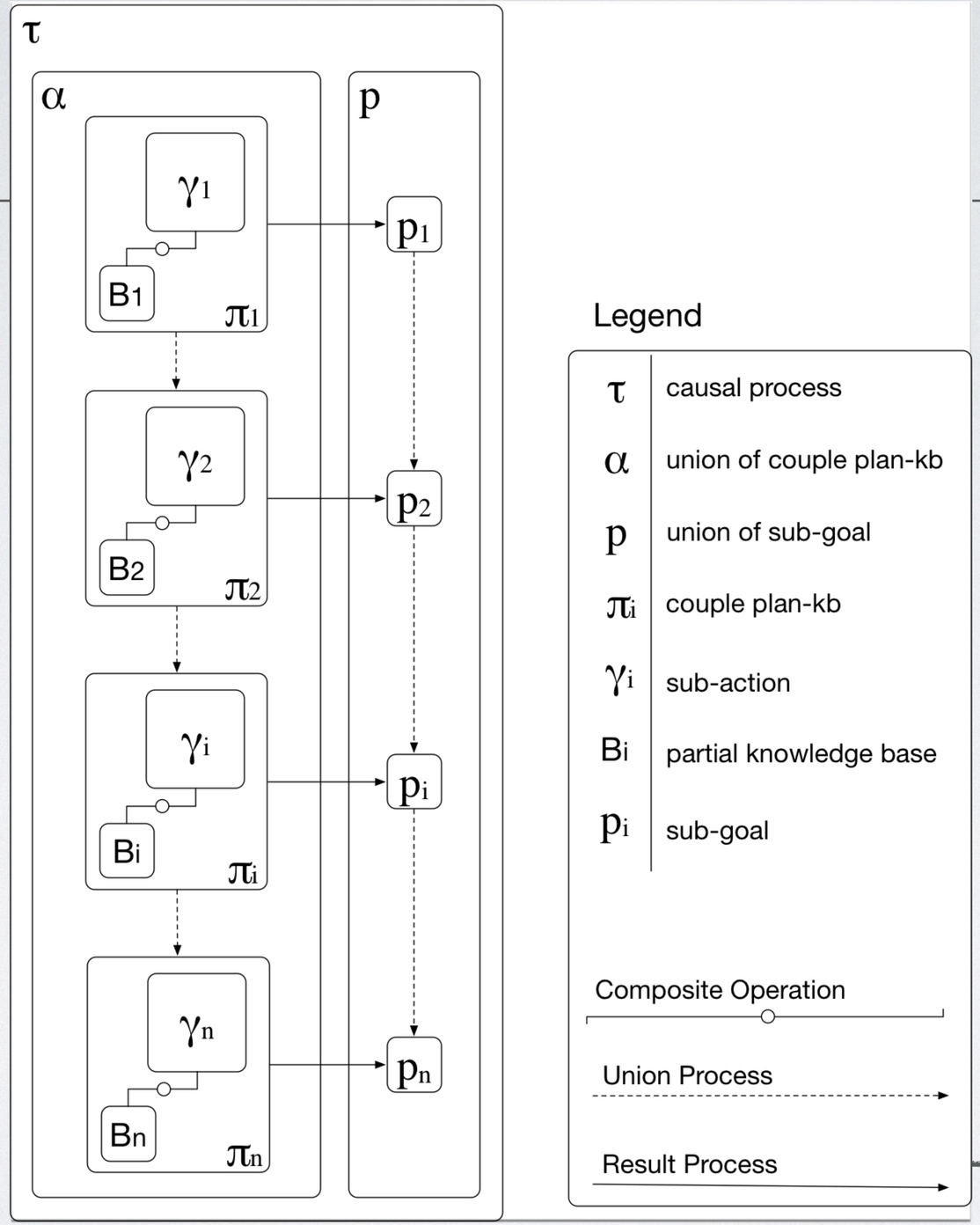
- Making beliefs explicit
- Breaking down actions and results —> plans and sub-results

$$TRUST(X\ Y\ C\ \tau\ g_X)$$

$$\text{where} \quad \tau = (\alpha, p) \quad \text{and} \quad g_X \equiv p;$$

$$\tau = (\alpha, p) \quad \text{where} \quad \alpha = \bigcup_{i=1}^{n} \pi_i \quad \text{and} \quad p = \bigcup_{i=1}^{n} p_i$$

$$\pi_i = \gamma_i \circ B_i \Rightarrow \alpha = \bigcup_{i=1}^{n} (\gamma_i \circ B_i)$$


τ
α
p
γ1
B1
π1
p1
γ2
B2
π2
p2
γi
Bi
πi
pi
γn
Bn
πn
pn
Legend
τ causal process
α union of couple plan-kb
p union of sub-goal
πi couple plan-kb
γi sub-action
Bi partial knowledge base
pi sub-goal
Composite Operation
Union Process
Result Process

```
1.  B ← B0;          /* B0 are initial beliefs */
2.  I ← I0;          /* I0 are initial intentions */
3.  while true do
4.    get next percept ρ via sensors;
5.    B ← brf(B, ρ);
6.    D ← options(B, I);
7.    I ← filter(B, D, I);
8.    π ← plan(B, I, Ac); /* Ac is the set of actions */
9.    while not (empty(π) or succeeded(I, B) or impossible(I, B)) do
10.     α ← first element of π;
11.     execute(α);
12.     π ← tail of π;
13.     observe environment to get next percept ρ;
14.     B ← brf(B, ρ);
15.     if reconsider(I, B) then
16.       D ← options(B, I);
17.       I ← filter(B, D, I);
18.     end-if
19.     if not sound(π, I, B) then
20.       π ← plan(B, I, Ac)
21.     end-if
22.   end-while
23. end-while
```

$$A_c \leftarrow action(B_{\alpha_i}, Cap)$$

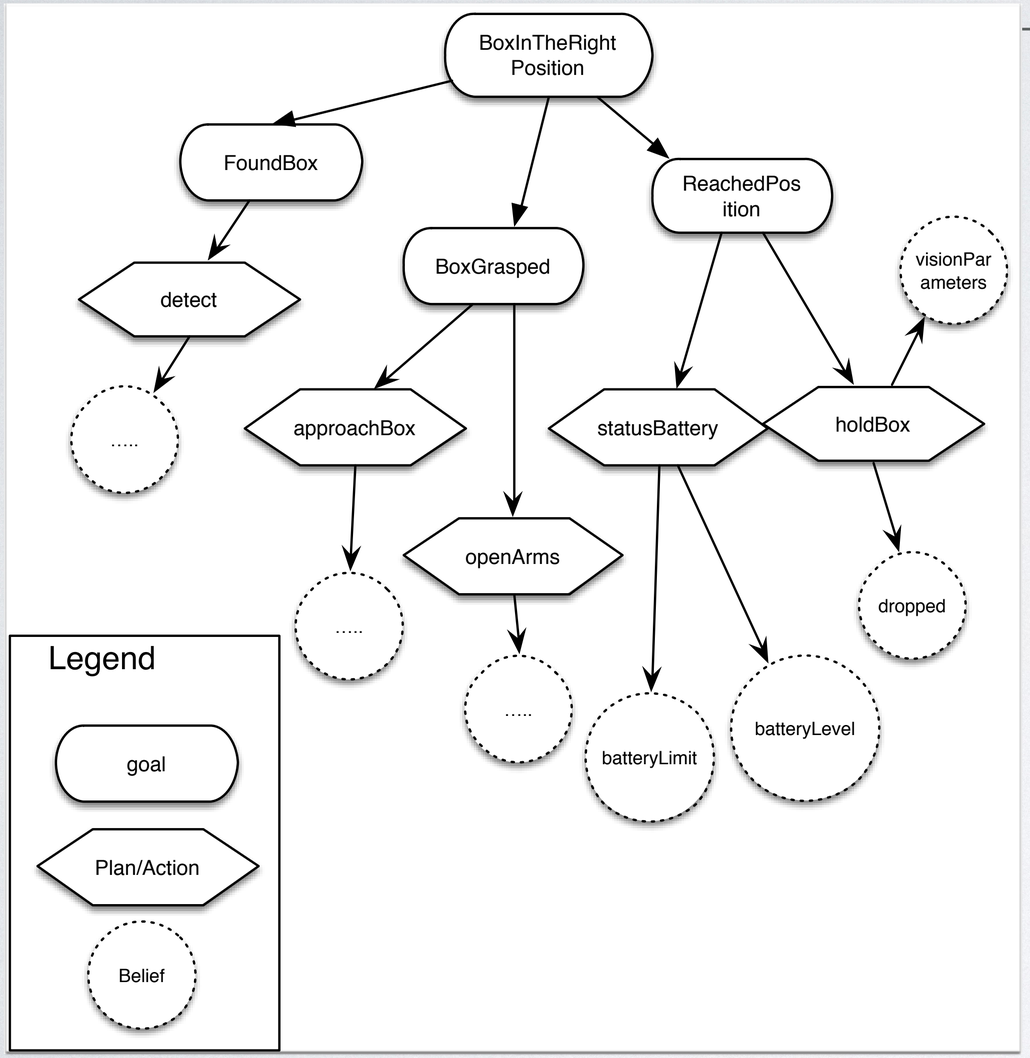
$$evaluate(\alpha_i);$$
$$J \leftarrow justify(\alpha_i, B_{\alpha_i});$$

Self-Modeling
Trustful interaction

THE ROBOT IN ACTION

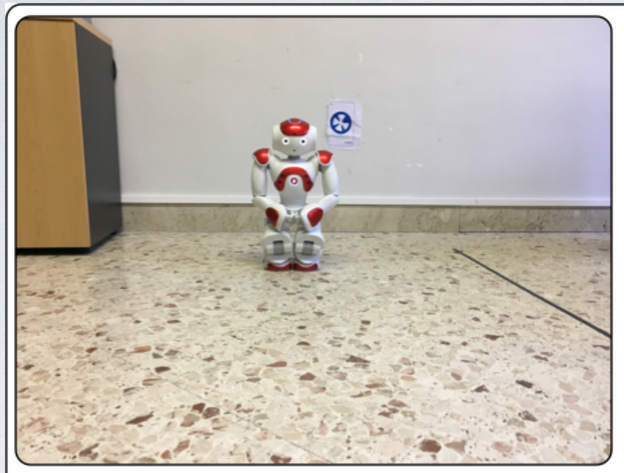
- Jason Agent
- CArtAgO artifact
- CArtAgO @Operation
- a reference model of the environment
  - all the internal elements of the agent/robot as part of the environment

THE ROBOT IN ACTION



1 +!ReachedPosition: true $\leftarrow$ goAhead; holdBox. $[\tau]$;
2 +!goAhead: batteryLimit(X) & batteryLevel(Y) & $Y < X \leftarrow$ say("My battery is exhaust. Please let me charge."). $[\gamma_1^+]$;
3 +!goAhead: batteryLimit(X) & batteryLevel(Y) & $Y \geq X \leftarrow$ execActions. $[\gamma_1^-]$;
4 B_1: batteryLimit, batteryLevel ;
5 +!holdBox: dropped(X) & visionParameters(Y) & $X == false \leftarrow$ execAct(Y). $[\gamma_2^+]$;
6 +!holdBox: dropped(X) & visionParameters(Y) & $X == true \leftarrow$ say("The box is dropped."). $[\gamma_2^-]$;
7 B_2: dropped, visionParameters ;

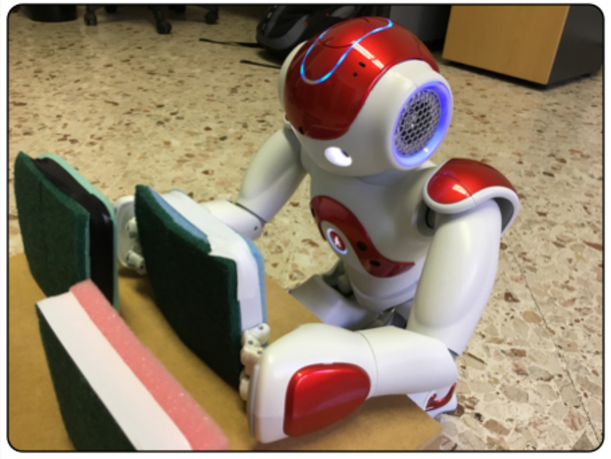
THE ROBOT IN ACTION



Initial Position

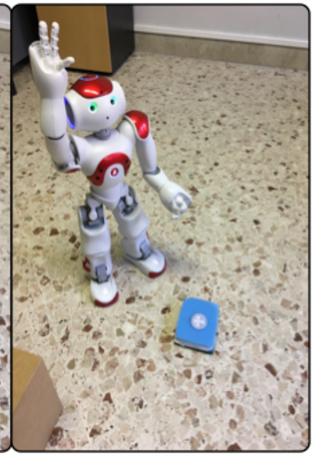
Boxes Assets

Detecting Phase

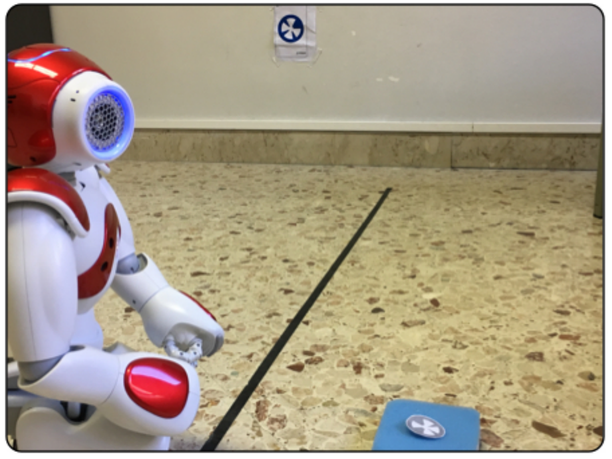


Taking Box

Box Dropped



Justification Phase



Box In The Right Position

CONCLUSIONS AND REMARKS

- Equipping the robot with self-modeling abilities
- Integrating Trust model with BDI deliberation process
- Exploiting JASON and CArtAgO
  - natively support BDI theory and have a well-established counterpart for actions, plans, knowledge implementation

CONCLUSIONS AND REMARKS

- Trust as a first element for triggering the decision process
- The integration in the BDI cycle —> two main elements of interaction
  - self-modeling
  - trust level in the interaction
- In the future:
  - implementing the other levels of adoption/delegation by Falcone&Castelfranchi
  - adding organization (MOISE)
  - adding other elements for triggering the decision process —> theory of mind

UNIVERSITÀ DEGLI STUDI DI PALERMO

dipartimento di ingegneria unipa

CNR ISTC INSTITUTE OF COGNITIVE SCIENCES AND TECHNOLOGIES

Thanks for your attention!

valeria.seidita@unipa.it