

This presentation is released under the terms of the  
**Creative Commons Attribution-Share Alike** license.

You are free to reuse it and modify it as much as you want as long as

- (1) you mention me as being the original author,
- (2) you re-share your presentation under the same terms.

You can download the sources of this presentation here:

[github.com/severin-lemaignan/presentation-iri-codesign-responsible-robotics](https://github.com/severin-lemaignan/presentation-iri-codesign-responsible-robotics)



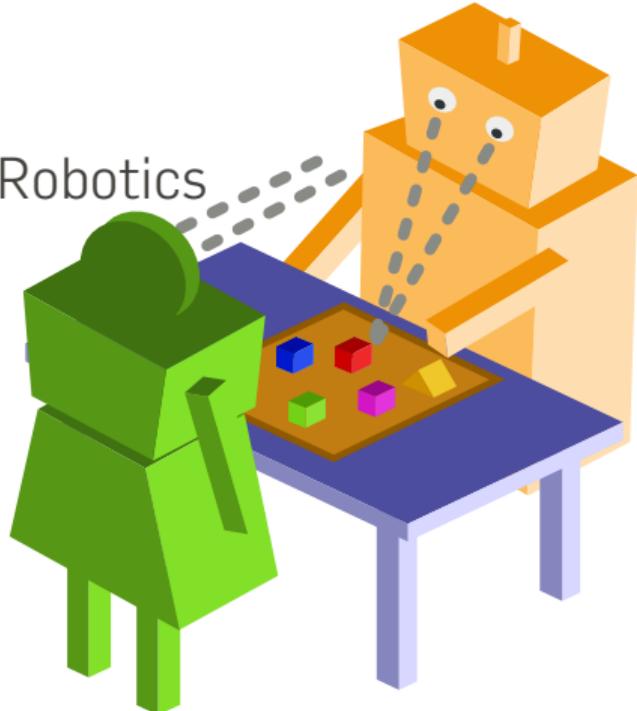
# End-to-end co-design and Responsible Robotics

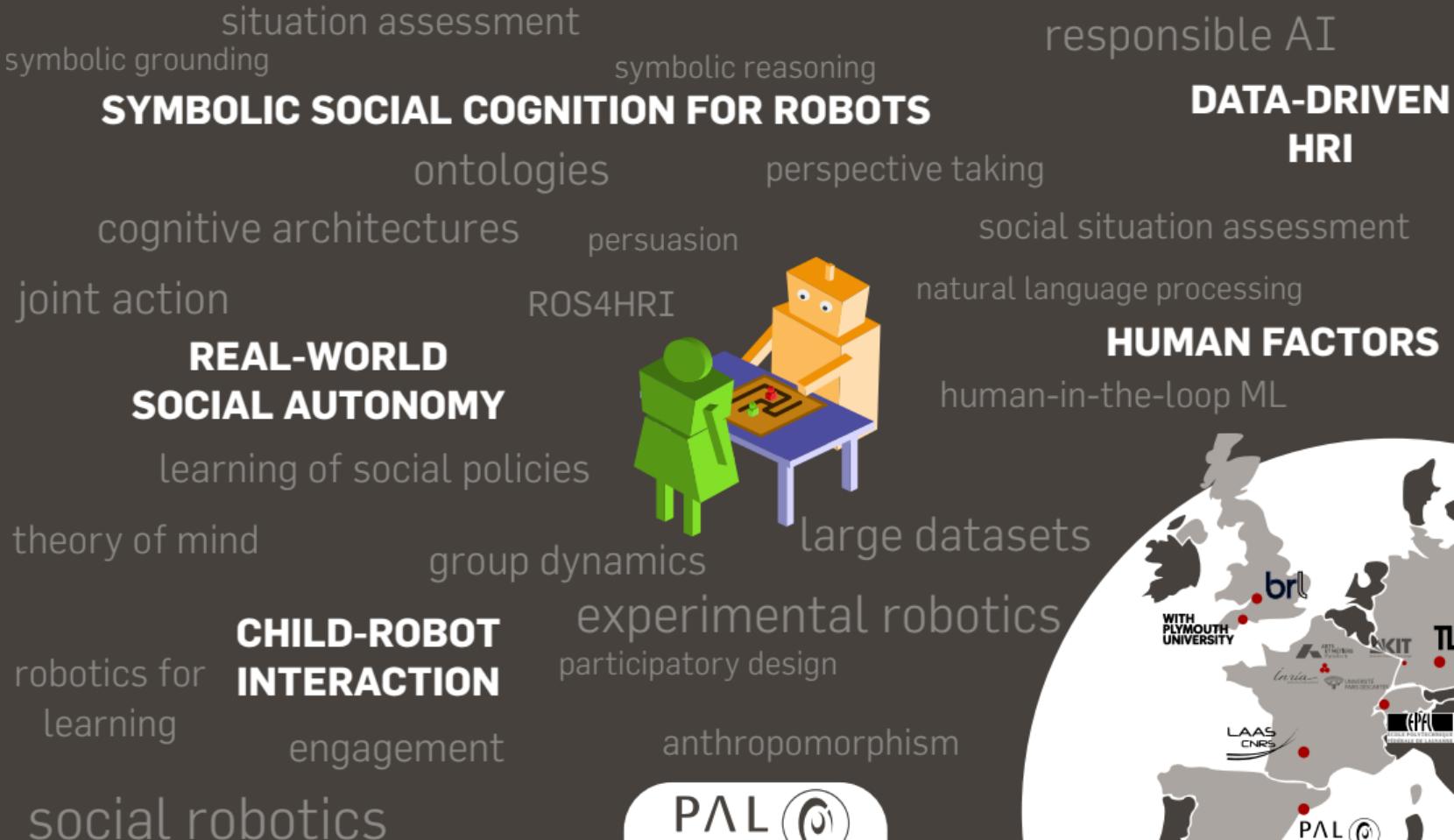
## Some food for thought in the context of HRI

Institut de Robòtica i Informàtica Industrial | UPC, 14 Dec 2023

Séverin Lemaignan

**PAL Robotics** Head of Social Robotics, Senior Scientist AI & HRI





# SOCIAL ROBOTICS

## Major scientific challenges:

- Model open-ended, underspecified situations; rich semantics; complex social dynamics;



# SOCIAL ROBOTICS

## Major scientific challenges:

- Model open-ended, underspecified situations; rich semantics; complex social dynamics;
- Close the interaction loop;



# SOCIAL ROBOTICS

## Major scientific challenges:

- Model open-ended, underspecified situations; rich semantics; complex social dynamics;
- Close the interaction loop;
- Understand and sustain long-term autonomous social interactions;



# SOCIAL ROBOTICS

## Major scientific challenges:

- Model open-ended, underspecified situations; rich semantics; complex social dynamics;
- Close the interaction loop;
- Understand and sustain long-term autonomous social interactions;
- Real-world algorithmic robustness;



# SOCIAL ROBOTICS

## Major scientific challenges:

- Model open-ended, underspecified situations; rich semantics; complex social dynamics;
- Close the interaction loop;
- Understand and sustain long-term autonomous social interactions;
- Real-world algorithmic robustness;
- Complex ethical landscape  
(correct, ethical = moving, context-based target);



# SOCIAL ROBOTICS

## Major scientific challenges:

- Model open-ended, underspecified situations; rich semantics; complex social dynamics;
- Close the interaction loop;
- Understand and sustain long-term autonomous social interactions;
- Real-world algorithmic robustness;
- Complex ethical landscape  
(correct, ethical = moving, context-based target);
- ⇒ cross-disciplinary & holistic approach required
- ⇒ involve all the stakeholders; participatory approach



## **Two questions I want to discuss today**

1. How far can we 'push' participatory design? Can we let the end-users do the hard work for us?

## **Two questions I want to discuss today**

1. How far can we 'push' participatory design? Can we let the end-users do the hard work for us?
2. If so, what are the consequences in terms of responsible robotics?

## LET SET OURSELVES A CHALLENGE

### **Design & run a study with:**

- a real robot

## LET SET OURSELVES A CHALLENGE

### Design & run a study with:

- a real robot
- a real interaction (...with a **human!**)

## LET SET OURSELVES A CHALLENGE

### Design & run a study with:

- a real robot
- a real interaction (...with a **human!**)
- a **continuous** interaction (several

## LET SET OURSELVES A CHALLENGE

### Design & run a study with:

- a real robot
- a real interaction (...with a **human!**)
- a **continuous** interaction (several)
- a **long** interaction (several weeks)

## LET SET OURSELVES A CHALLENGE

### Design & run a study with:

- a real robot
- a real interaction (...with a **human!**)
- a **continuous** interaction (several)
- a **long** interaction (several weeks)
- a realistic task (**large state vector & action space**)

## LET SET OURSELVES A CHALLENGE

### Design & run a study with:

- a real robot
- a real interaction (...with a **human!**)
- a **continuous** interaction (several)
- a **long** interaction (several weeks)
- a realistic task (**large state vector & action space**)
- **in the wild**

## LET SET OURSELVES A CHALLENGE

### Design & run a study with:

- a real robot
- a real interaction (...with a **human!**)
- a **continuous** interaction (several)
- a **long** interaction (several weeks)
- a realistic task (**large state vector & action space**)
- **in the wild**
- also including **social behaviours & social dynamics**

# LET SET OURSELVES A CHALLENGE

## Design & run a study with:

- a real robot
- a real interaction (...with a **human!**)
- a **continuous** interaction (several)
- a **long** interaction (several weeks)
- a realistic task (**large state vector & action space**)
- **in the wild**
- also including **social behaviours & social dynamics**
- ...and of course, the robot should be **autonomous**

---

## 'END-TO-END' CO-DESIGN



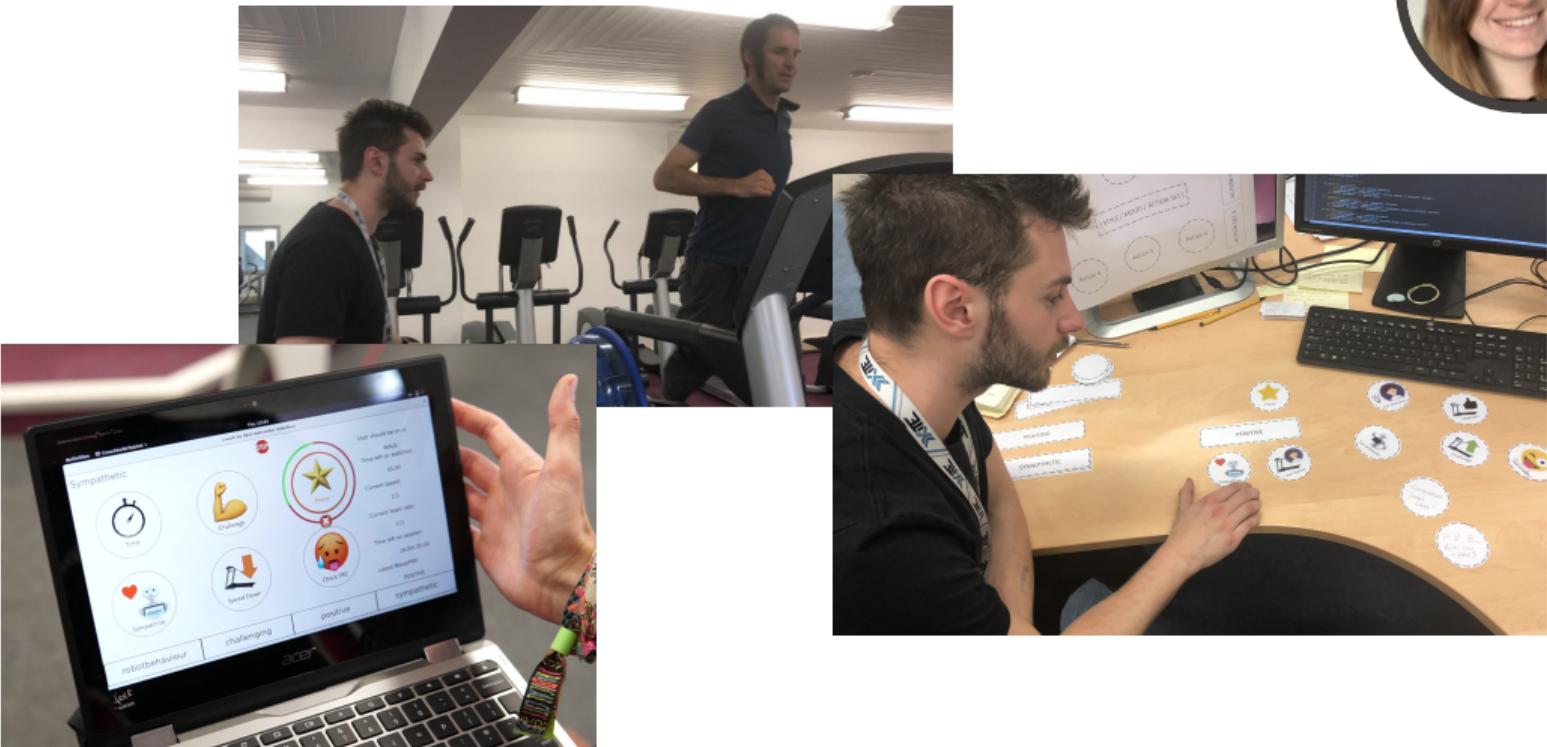
End-to-end co-design: gym coach

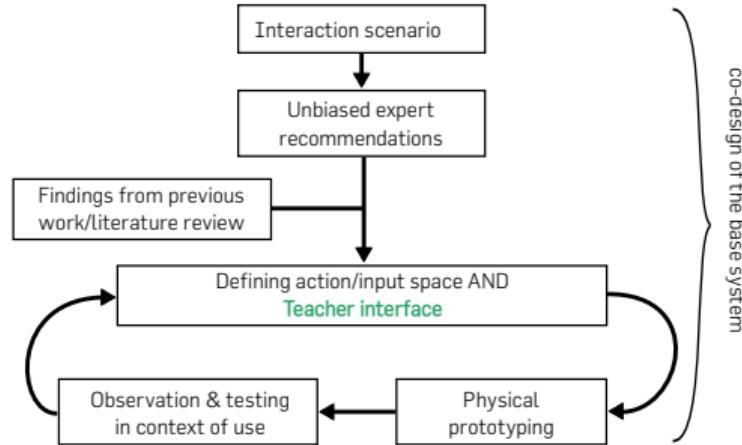
## COUCH-TO-5K STUDY

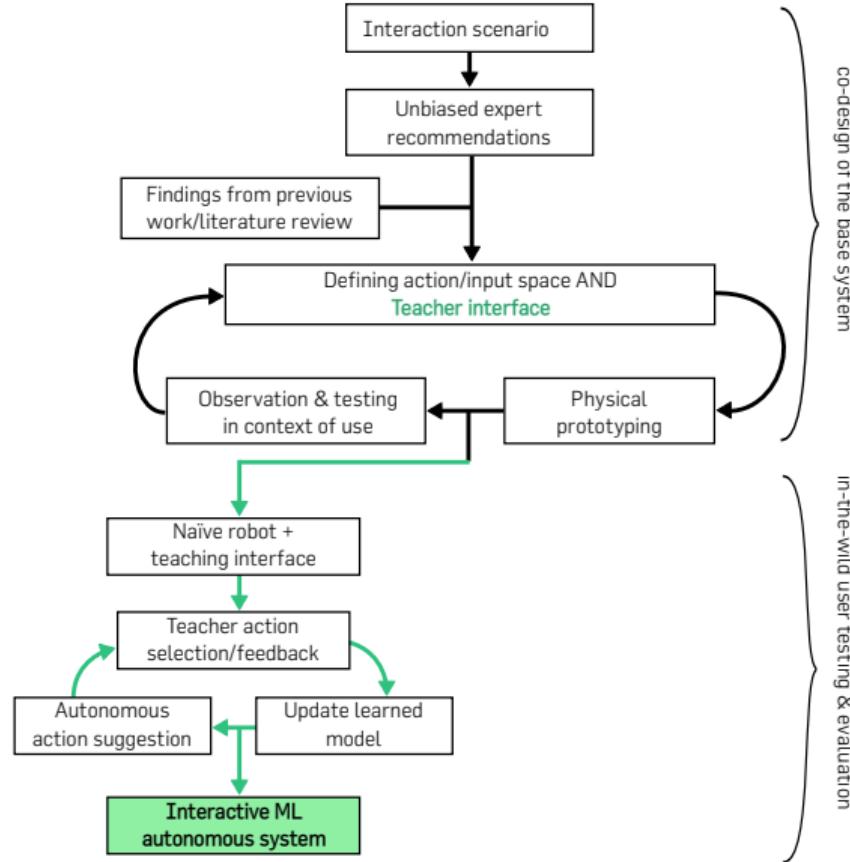


- 9 participants
- 3 months; 27 one-hour sessions per participants
- $|state| = 20$ ;  $|action\_space| = 11$
- Includes participants' personality (Big-5) as input feature

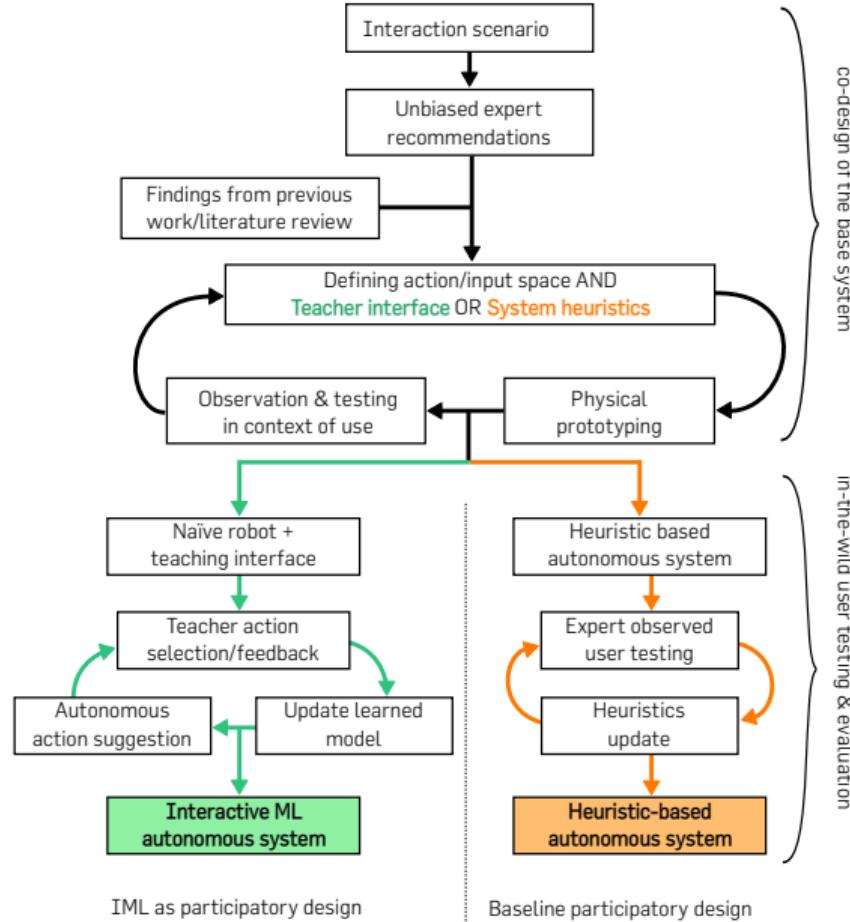
# CO-DESIGN FOR REAL-WORLD + LONG-TERM



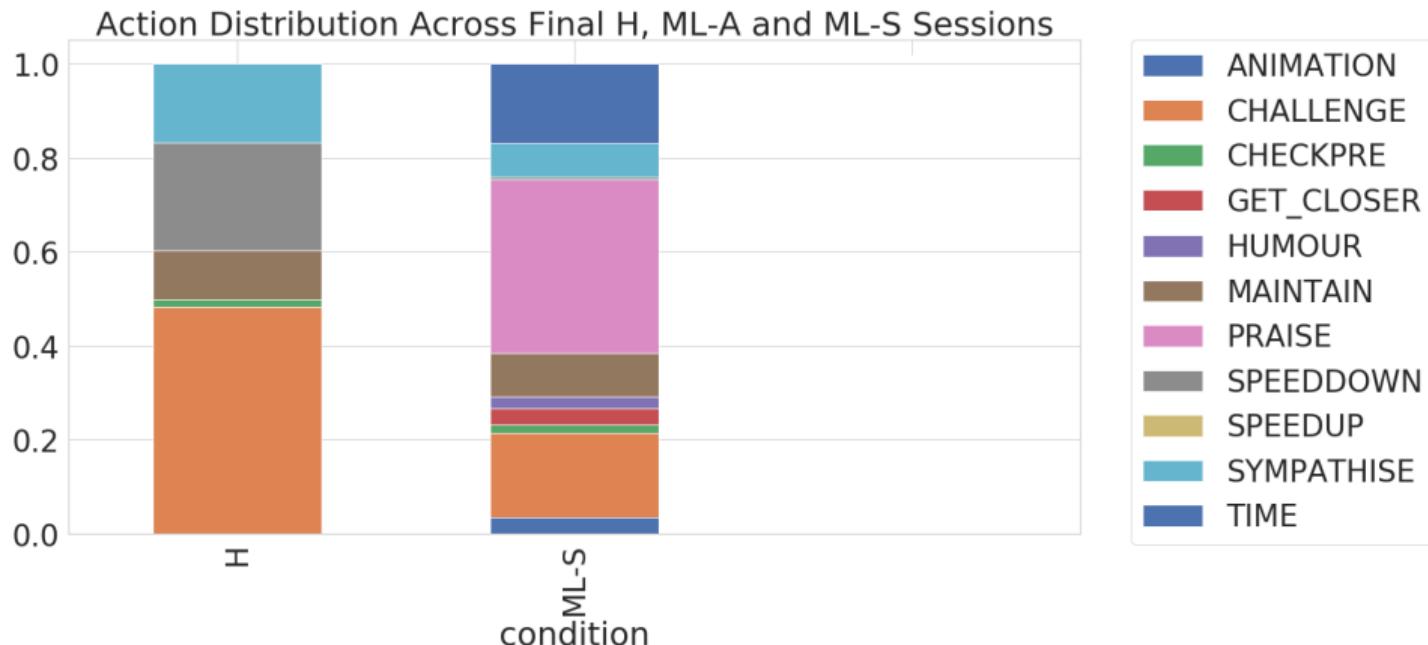




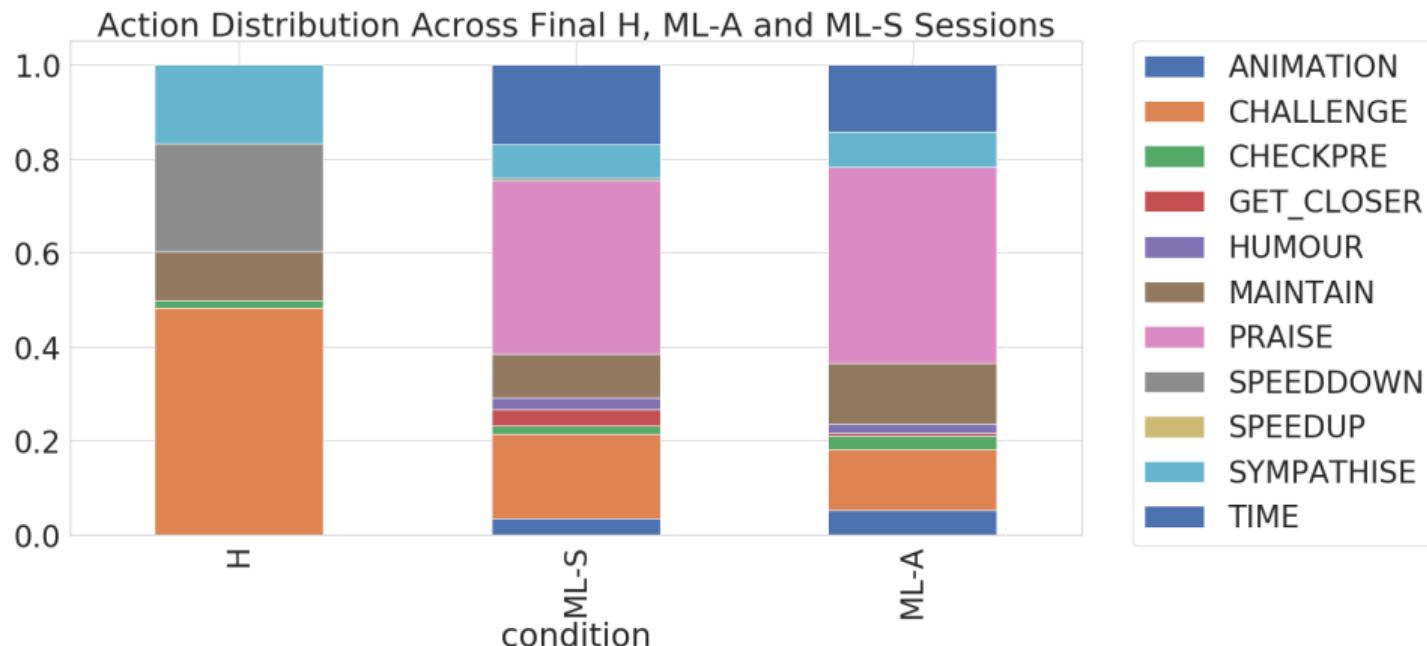
IML as participatory design



# LEARNT POLICIES

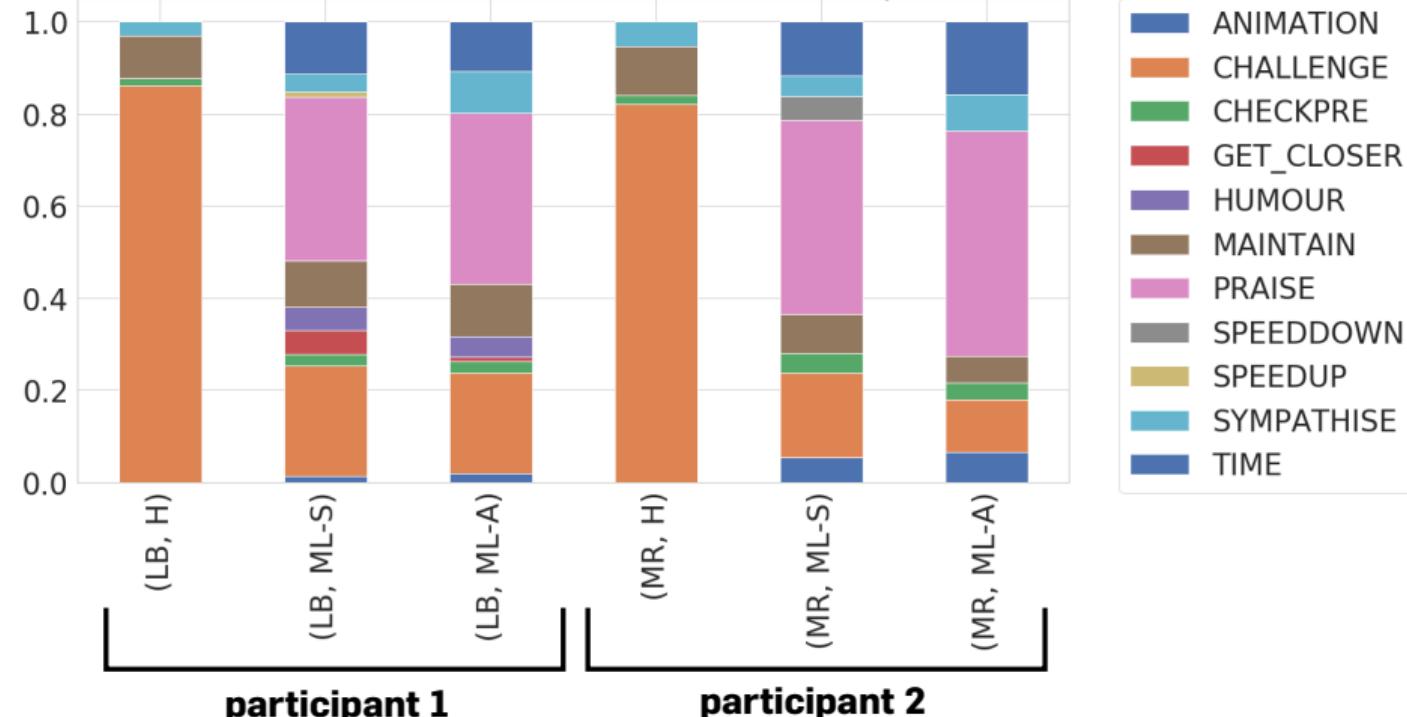


# LEARNT POLICIES



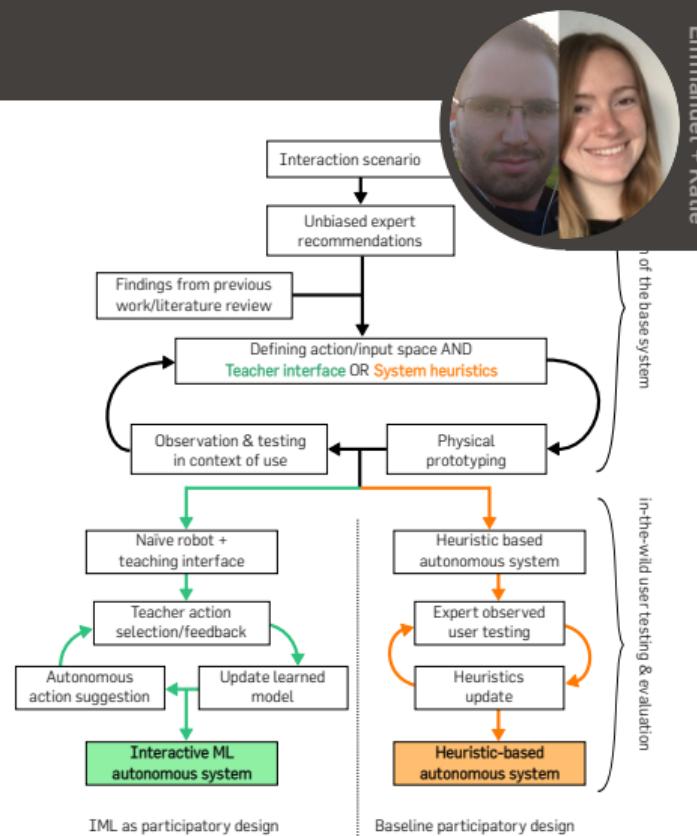
# LEARNT POLICIES

Phase 3 H, ML-A and ML-S Action Distribution for Participants LB and MR



# LEADOR: END-TO-END CO-DESIGN

- o a successful technical solution to replace the wizard;
- o but equally important, a **end-to-end** participatory design methodology (*LEADOR: Led-by-Experts Automation and Design Of Robots*)



More importantly,

“It was like training my assistant. Now I let it start with the next participant, while I do stretching with the previous one.”

---

# RESPONSIBLE ROBOTICS

## EU GUIDELINES FOR TRUSTWORTHY AI CHECKLIST

- human autonomy, human agency
- consent
- human oversight, especially if the system evolves (ML)
- technical robustness and safety
- respect for privacy, data ownership
- traceability, explainability, communication
- unfair bias avoidance, stakeholder participation
- sustainability
- social impact
- Accountability, documenting trade-offs



# RESPONSIBLE ROBOTICS AND XAI

In EU's Ethics Guidelines for Trustworthy AI, keywords like:

- transparency
- explainability (eg XAI)
- traceability (eg Winfield's *ethical black box*; TRAIL project)

## RESPONSIBLE ROBOTICS AND XAI

In EU's Ethics Guidelines for Trustworthy AI, keywords like:

- transparency
- explainability (eg XAI)
- traceability (eg Winfield's *ethical black box*; TRAIL project)

Useful mechanistic approach, ...but does it capture well the *feeling* of 'understanding' the machine? (ie, in-situ, in-the-moment, intuitive understanding)

# RESPONSIBLE ROBOTICS AND XAI



In EU's Ethics Guidelines for Trustworthy AI, keywords like:

- transparency
- explainability (eg XAI)
- tracability (eg Winfield's *ethical black box*; TRAIL project)

Useful mechanistic approach, ...but does it capture well the *feeling* of 'understanding' the machine? (ie, in-situ, in-the-moment, intuitive understanding)

In other words, would eg explainability and tracability by themselves be sufficient to establish trust? (trust *in bridge* vs trust *in people*?)

⇒ **understandability ?**

## DO WE ACTUALLY REALLY WANT TO UNDERSTAND OUR ROBOTS?

Users (ourselves included) want technology that 'just work'; cf Obadia's idea of the inherent *magic* of technology

## DO WE ACTUALLY REALLY WANT TO UNDERSTAND OUR ROBOTS?

Users (ourselves included) want technology that 'just work'; cf Obadia's idea of the inherent *magic* of technology

Example: when people ask something to ChatGPT, they might complain that the result is incorrect (hallucinated or otherwise) ⇒ they want the visible, *surface* result to fixed.

## DO WE ACTUALLY REALLY WANT TO UNDERSTAND OUR ROBOTS?

Users (ourselves included) want technology that 'just work'; cf Obadia's idea of the inherent *magic* of technology

Example: when people ask something to ChatGPT, they might complain that the result is incorrect (hallucinated or otherwise) ⇒ they want the visible, *surface* result to fixed.

Do they want to *understand* why ChatGPT got it wrong? I'd argue that the vast majority do not really care. Also, psychological angle, eg Nass' Media Equation

# DO WE ACTUALLY REALLY WANT TO UNDERSTAND OUR ROBOTS?



Users (ourselves included) want technology that 'just work'; cf Obadia's idea of the inherent *magic* of technology

Example: when people ask something to ChatGPT, they might complain that the result is incorrect (hallucinated or otherwise) ⇒ they want the visible, *surface* result to fixed.

Do they want to *understand* why ChatGPT got it wrong? I'd argue that the vast majority do not really care. Also, psychological angle, eg Nass' Media Equation

However, I would also argue that, to trust the system, what we really want is to be able to confidently build a robust mental model , with a strong predictive power.

# DO WE ACTUALLY REALLY WANT TO UNDERSTAND OUR ROBOTS?



Users (ourselves included) want technology that 'just work'; cf Obadia's idea of the inherent *magic* of technology

Example: when people ask something to ChatGPT, they might complain that the result is incorrect (hallucinated or otherwise) ⇒ they want the visible, *surface* result to fixed.

Do they want to *understand* why ChatGPT got it wrong? I'd argue that the vast majority do not really care. Also, psychological angle, eg Nass' Media Equation

However, I would also argue that, to trust the system, what we really want is to be able to **confidently** build a robust mental model , with a strong predictive power.

# DO WE ACTUALLY REALLY WANT TO UNDERSTAND OUR ROBOTS?



Users (ourselves included) want technology that 'just work'; cf Obadia's idea of the inherent *magic* of technology

Example: when people ask something to ChatGPT, they might complain that the result is incorrect (hallucinated or otherwise) ⇒ they want the visible, *surface* result to fixed.

Do they want to *understand* why ChatGPT got it wrong? I'd argue that the vast majority do not really care. Also, psychological angle, eg Nass' Media Equation

However, I would also argue that, to trust the system, what we really want is to be able to confidently build a **robust mental model**, with a strong predictive power.

# DO WE ACTUALLY REALLY WANT TO UNDERSTAND OUR ROBOTS?



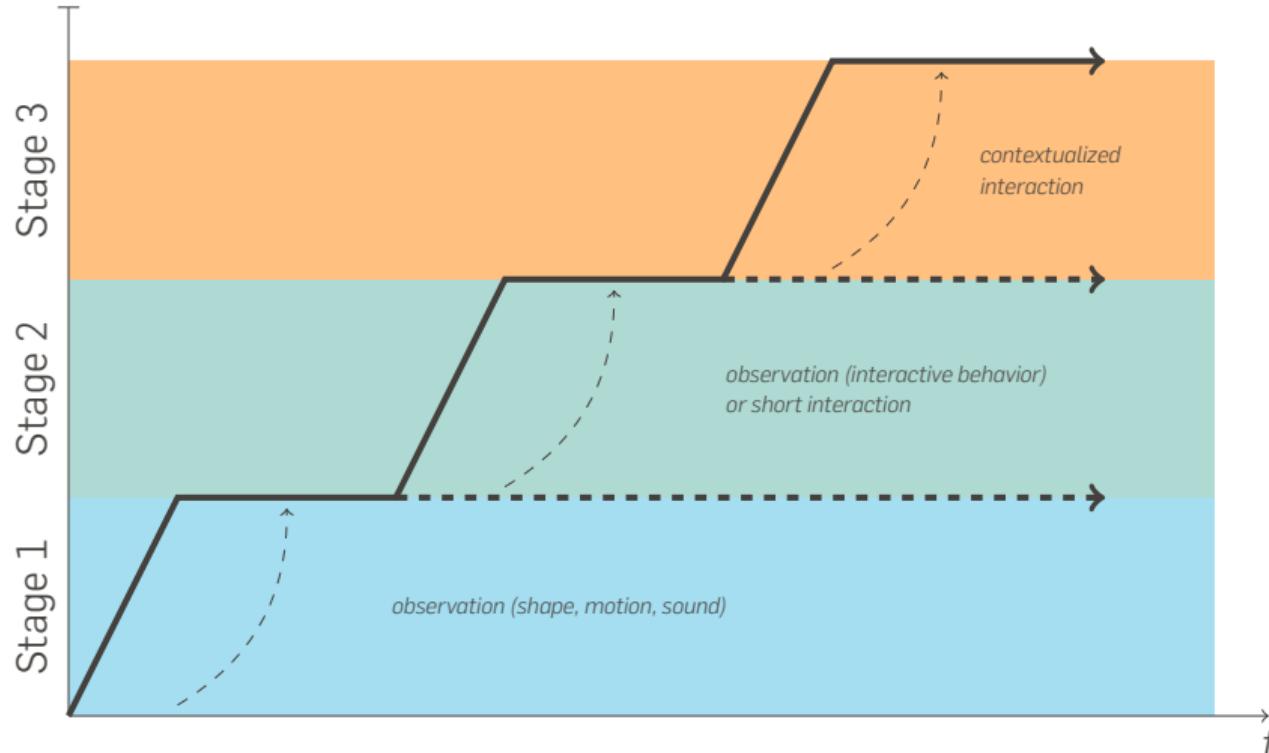
Users (ourselves included) want technology that 'just work'; cf Obadia's idea of the inherent *magic* of technology

Example: when people ask something to ChatGPT, they might complain that the result is incorrect (hallucinated or otherwise) ⇒ they want the visible, *surface* result to fixed.

Do they want to *understand* why ChatGPT got it wrong? I'd argue that the vast majority do not really care. Also, psychological angle, eg Nass' Media Equation

However, I would also argue that, to trust the system, what we really want is to be able to confidently build a robust mental model , with a strong **predictive power**.

# COGNITIVE INTERPRETATION?



adapted mental model  
→ adapted interaction  
(anthropomorphic or not)

projection of existing  
mental models  
(familiarity)

pre-cognitive  
anthropomorphism



Julia Fink

## WHAT MEANS 'RESPONSIBLE' IN THAT CONTEXT?

If we adopt the cognitive perspective of *understanding* the robot as the **completion of a robust mental model** of the machine, where does that leave us wrt. *responsible understandability*?

## WHAT MEANS 'RESPONSIBLE' IN THAT CONTEXT?

If we adopt the cognitive perspective of *understanding* the robot as the **completion of a robust mental model** of the machine, where does that leave us wrt. *responsible understandability*?

⇒ design robots' behaviours to foster the creation of detailed and correct mental models

## WHAT MEANS 'RESPONSIBLE' IN THAT CONTEXT?



If we adopt the cognitive perspective of *understanding* the robot as the **completion of a robust mental model** of the machine, where does that leave us wrt. *responsible understandability*?

⇒ design robots' behaviours to foster the creation of detailed and correct mental models

...or even let the end-users co-design the robot's cognitive capabilities



“It was like training my assistant. Now I let it start with the next participant, while I do stretching with the previous one.”

...does he *understand* the robot? yes and no

...is it *Responsible Robotics*©?

---

## BRIEF LOOK AT THE INDUSTRY

## DOES IT TRANSLATE TO INDUSTRY?

Can we turn the LEADOR paradigm into a generic & robust enough technique for broad usage?



# DOES IT TRANSLATE TO INDUSTRY?

Can we turn the LEADOR paradigm into a generic & robust enough technique for broad usage?

## Two key use-cases for PAL Robotics:

- robots that know when to help in public administrations
- elderly care and isolation (but... **rHHI**: robot-supported Human-Human Interactions!)



SOCIAL CONTACT



## ACTIVITIES



CORE ACTIVITIES



EMERGENCY DEVICES



## TECHNOLOGY RELATED



## PERSONALITIES



## SELF VIEW



## FEARS



### PROBLEM

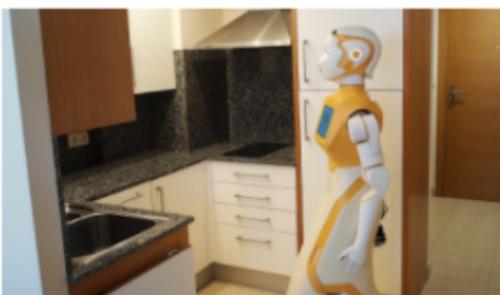


# NHoA: Never HOMe Alone

## Clinica Humana



Older person home (reception)



Can Granada sheltered apartment

H2020 SHAPES



Older person home (terrace)

SALIDA DE  
EMERGENCIA

AMIBA



AMIBA



AMIBA

## DOES IT TRANSLATE TO INDUSTRY?

- practical issues are typically overlooked (co-design in-situ)



## DOES IT TRANSLATE TO INDUSTRY?

- practical issues are typically overlooked (co-design in-situ)
- adoption: primarily from *secondary* users (carers, teachers...)



## DOES IT TRANSLATE TO INDUSTRY?

- practical issues are typically overlooked (co-design in-situ)
- adoption: primarily from *secondary* users (carers, teachers...)
- needs are not always what we think (co-design with all stakeholders)



## DOES IT TRANSLATE TO INDUSTRY?

- practical issues are typically overlooked (co-design in-situ)
- adoption: primarily from *secondary* users (carers, teachers...)
- needs are not always what we think (co-design with all stakeholders)
- robots do not live in isolation: think 'eco-systems' and embrace "mis-use"/mis-understanding of the robot



## DOES IT TRANSLATE TO INDUSTRY?

- practical issues are typically overlooked (co-design in-situ)
- adoption: primarily from *secondary users* (carers, teachers...)
- needs are not always what we think (co-design with all stakeholders)
- robots do not live in isolation: think 'eco-systems' and embrace "mis-use"/mis-understanding of the robot
- from interaction to habits: (individualised) long-term *mutual shaping*



## DOES IT TRANSLATE TO INDUSTRY?

- practical issues are typically overlooked (co-design in-situ)
- adoption: primarily from *secondary users* (carers, teachers...)
- needs are not always what we think (co-design with all stakeholders)
- robots do not live in isolation: think 'eco-systems' and embrace "mis-use"/mis-understanding of the robot
- from interaction to habits: (individualised) long-term *mutual shaping*
- importance of the regulatory framework (legal + ethical) to steer an otherwise 'engineering-focused' mindset





## SOME FINAL REMARKS

Co-design requires bootstrapping: you need to understand at *some* level the physical (easy) and cognitive (much harder) capabilities of the robot to 'invent' its role

→ iterative process

→ slow process

(though more of a problem for traditional co-design than lifelong, end-to-end co-design)

## SOME FINAL REMARKS

Sabanovic's *mutual shaping* is fundamental and not yet fully internalised by researchers/engineers

→ you (re)shape your relation to – and understanding of – the robot by using it

⇒ *Understanding* the robot likely means that the robot also has to understand you!

(here, the meaning of *understand* is left as an exercise to the reader)

## SOME FINAL REMARKS

...as such:

can we really separate the human's understanding of the robot, from the robot's understanding of the human, of its environment, and... of itself? (introspection)

## ANECDOTE: WHAT IF WE ASK THE SYSTEM TO EXPLAIN ITS BEHAVIOUR?

Sally (a girl) has 3 brothers. Each brother has 2 sisters. How many sisters does Sally have?

# ANECDOTE: WHAT IF WE ASK THE SYSTEM TO EXPLAIN ITS BEHAVIOUR



Sally (a girl) has 3 brothers. Each brother has 2 sisters. How many sisters does Sally have?

GPT-4:

Sally has 2 sisters.

## ANECDOTE: WHAT IF WE ASK THE SYSTEM TO EXPLAIN ITS BEHAVIOUR



Sally (a girl) has 3 brothers. Each brother has 2 sisters. How many sisters does Sally have? Let's think step by step.

# ANECDOTE: WHAT IF WE ASK THE SYSTEM TO EXPLAIN ITS BEHAVIOUR



Sally (a girl) has 3 brothers. Each brother has 2 sisters. How many sisters does Sally have? Let's think step by step.

GPT-4:

Sally has 3 brothers. Each of these brothers has 2 sisters. This means that there are 2 girls in the family, including Sally. Therefore, Sally has 1 sister.

(incidentally, GPT-4 is the only model out of 60 tested that get the correct answer)

## TAKE-HOME MESSAGES

- **IML for 'end-to-end' participatory design:** it works! requires however a well-understood domain

## TAKE-HOME MESSAGES

- **IML for 'end-to-end' participatory design:** it works! requires however a well-understood domain
- **to make them accept/trust the robot, make them understand it!**: *understand* has however a very different meaning for the regulator/designer, and for the non-expert end-user. **Analytic understanding != intuitive understanding**, ie cognitive modeling

## TAKE-HOME MESSAGES

- **IML for 'end-to-end' participatory design:** it works! requires however a well-understood domain
- **to make them accept/trust the robot, make them understand it!**: *understand* has however a very different meaning for the regulator/designer, and for the non-expert end-user. **Analytic understanding != intuitive understanding**, ie cognitive modeling
- **robots exist in whole ecosystems:** need to stop thinking of HRI as dyadic interactions; multiple stakeholders, potentially changing stakeholders, with wildly different needs and goals

## TAKE-HOME MESSAGES

- **IML for 'end-to-end' participatory design:** it works! requires however a well-understood domain
- **to make them accept/trust the robot, make them understand it!**: *understand* has however a very different meaning for the regulator/designer, and for the non-expert end-user. **Analytic understanding != intuitive understanding**, ie cognitive modeling
- **robots exist in whole ecosystems:** need to stop thinking of HRI as dyadic interactions; multiple stakeholders, potentially changing stakeholders, with wildly different needs and goals
- **lifelong responsibility on shifting sands:** interaction is reshaped over time by usage, accommodation, individual personality traits; responsibility does not stop at the point of sale



# Thank you!

We are always looking for  
great people to join us!

Drop me a line if you want to know more



KNOW



**PAL**  
ROBOTICS



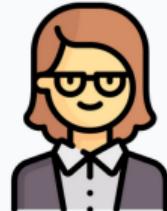
robot designer  
+ manufacturer



field technicians



retailer



manager

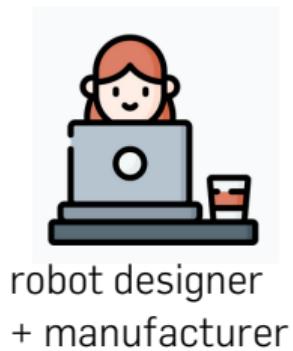
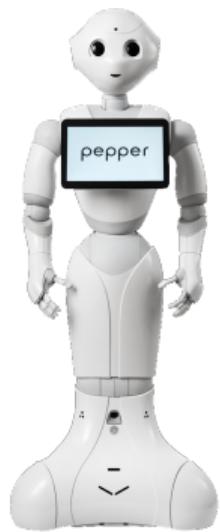


employees



customers (?)





robot designer  
+ manufacturer



manager



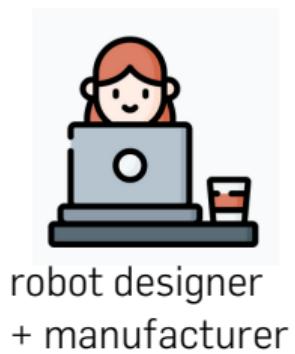
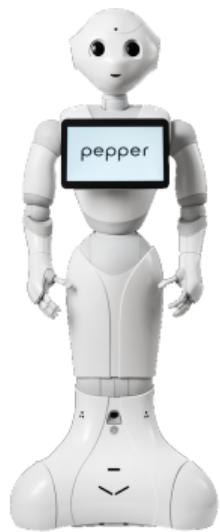
gym



coach



customers



robot designer  
+ manufacturer



manager



gym



power



coach



customers



power





regulator

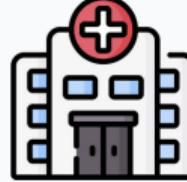
PAL  
ROBOTICS



robot designer  
+ manufacturer



integrator



hospital



nurses



patients



manager



families