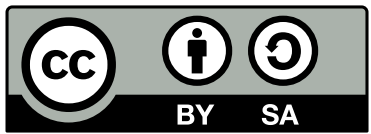


This presentation is released under the terms of the
Creative Commons Attribution-Share Alike license.

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/2023-understandability-dagstuhl

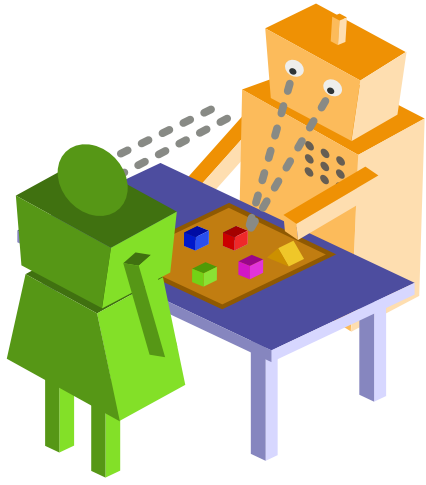
PAL ROBOTICS

ACCIÓ Generalitat de Catalunya

Understandability

Some food for thought in the context of HRI

Responsible Robotics | Dagstuhl Seminar, 11-15 Sep 2023



Séverin Lemaignan

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

situation assessment

symbolic grounding

symbolic reasoning

responsible AI

SYMBOLIC SOCIAL COGNITION FOR ROBOTS

DATA-DRIVEN HRI

ontologies

perspective taking

cognitive architectures

persuasion

social situation assessment

joint action

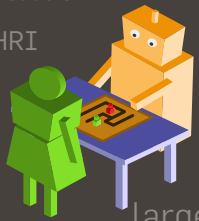
ROS4HRI

natural language processing

REAL-WORLD SOCIAL AUTONOMY

HUMAN FACTORS

human-in-the-loop ML



learning of social policies

theory of mind

group dynamics

large datasets

experimental robotics

CHILD-ROBOT INTERACTION

participatory design

robotics for learning

engagement

anthropomorphism

social robotics

PAL ROBOTICS

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;

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;
- $\Rightarrow$ cross-disciplinary & holistic approach required
- $\Rightarrow$ involve all the stakeholders; participatory approach

UNDERSTANDABILITY

…‘Understandability’?

UNDERSTANDABILITY

…‘Understandability’?

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

- transparency
- explainability (eg XAI)
- tracability (eg Winfield’s *ethical black box*)

UNDERSTANDABILITY

…‘Understandability’?

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

- transparency
- explainability (eg XAI)
- tracability (eg Winfield’s *ethical black box*)

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

UNDERSTANDABILITY

…‘Understandability’?

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

- transparency
- explainability (eg XAI)
- tracability (eg Winfield’s *ethical black box*)

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?

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) $\Rightarrow$ 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) $\Rightarrow$ 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.

DO WE ACTUALLY REALLY WANT TO UNDERSTAND OUR ROBOTS?

PAL ROBOTICS

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) $\Rightarrow$ 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.

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?

PAL ROBOTICS

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) $\Rightarrow$ 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.

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) $\Rightarrow$ 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.

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?

PAL ROBOTICS

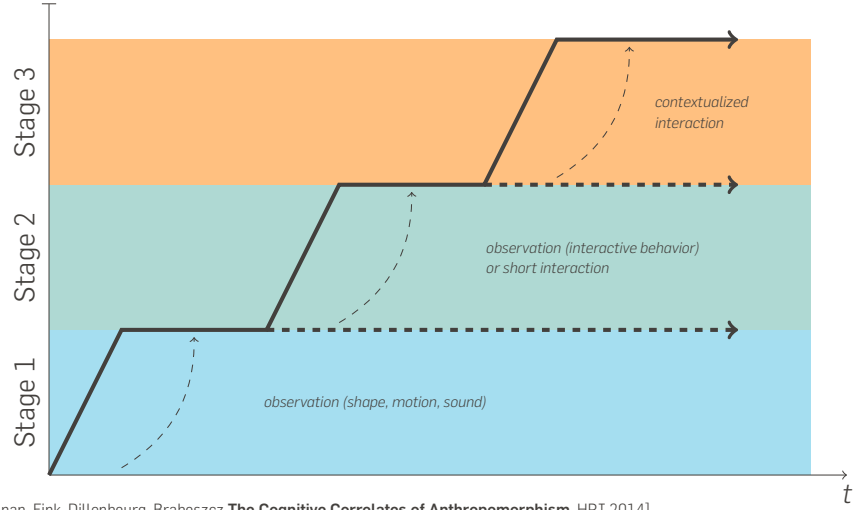
Users (ourselves included) want technology that 'just work'; cf Obadia's idea of the inherent *magic* of technology

Example: when people ask something to ChatGPT, they might complain that the result is incorrect (hallucinated or otherwise) $\Rightarrow$ 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.

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

[Lemaignan, Fink, Dillenbourg, Braboszcz **The Cognitive Correlates of Anthropomorphism**, HRI 2014]

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?

$\Rightarrow$ design robots’ behaviours to foster the creation of detailed and correct mental models

WHAT MEANS ‘RESPONSIBLE’ IN THAT CONTEXT?

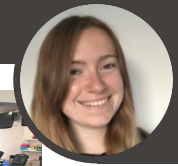
If we adopt the cognitive perspective of *understanding* the robot as the **completion of a robust mental model** of the machine, where does that leave us wrt. *responsible* understandability?

$\Rightarrow$ 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

End-to-end co-design: gym coach

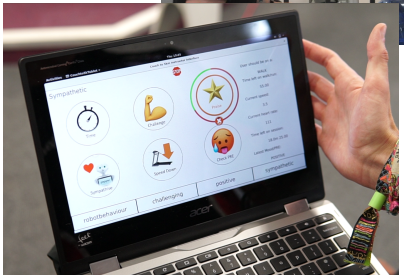
COUCH-TO-5K STUDY



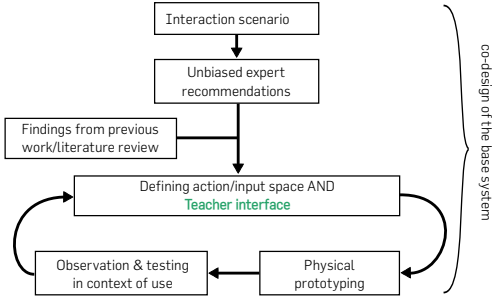
- 9 participants
- 3 months; 27 one-hour sessions per participants
- $|state| = 20$; $|action\_space| = 11$
- Includes participants' personality (Big-5) as input feature

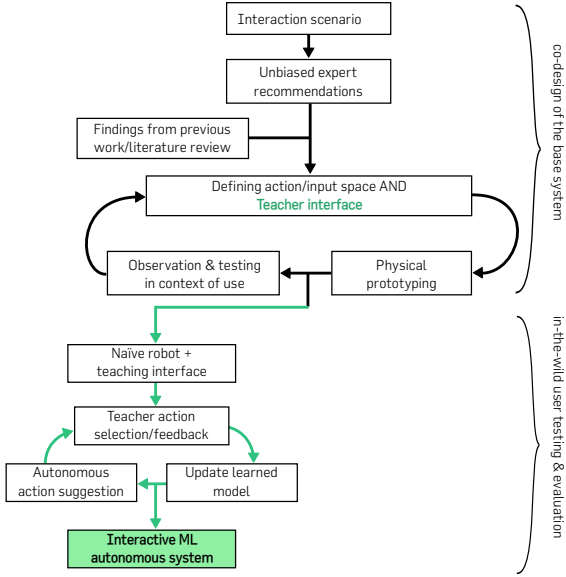
[Winkle et al. **In-Situ Learning from a Domain Expert for Real World Socially Assistive Robot Deployment** RSS 2020]

CO-DESIGN FOR REAL-WORLD + LONG-TERM

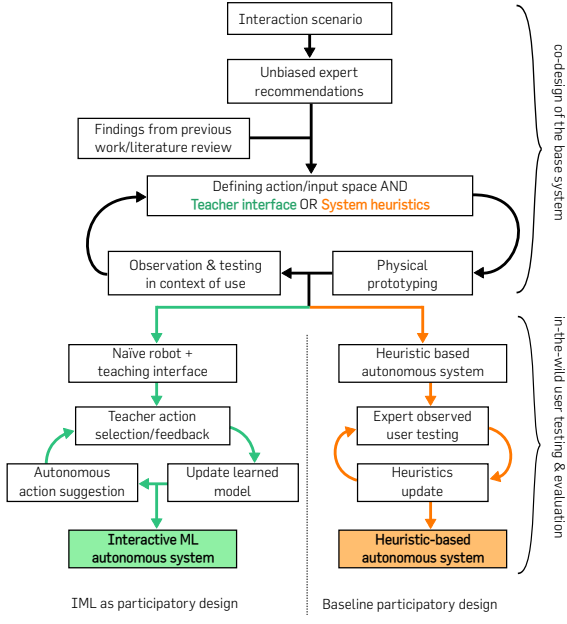


[Winkle et al. **In-Situ Learning from a Domain Expert for Real World Socially Assistive Robot Deployment** RSS 2020]

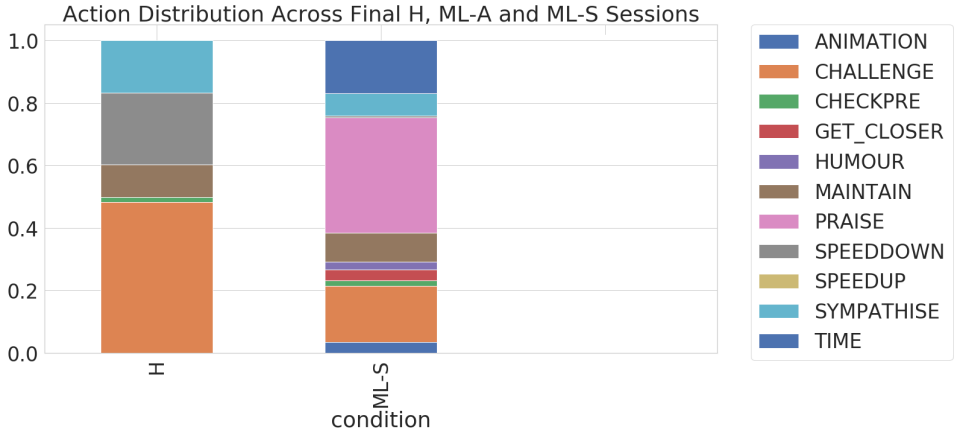

Interaction scenario
Unbiased expert recommendations
Findings from previous work/literature review
Defining action/input space AND
Teacher interface
Observation & testing in context of use
Physical prototyping
co-design of the base system



IML as participatory design

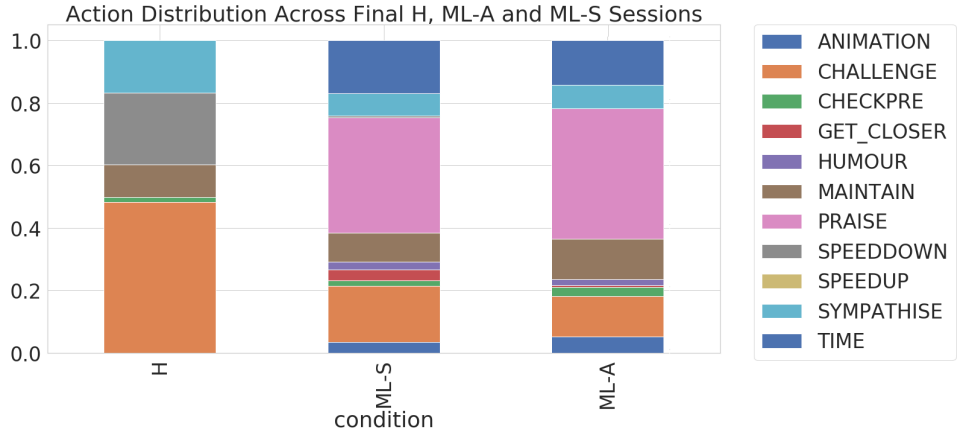

Interaction scenario
Unbiased expert recommendations
Findings from previous work/literature review
Defining action/input space AND
Teacher interface OR System heuristics
Observation & testing in context of use
Physical prototyping
co-design of the base system
Naïve robot + teaching interface
Teacher action selection/feedback
Autonomous action suggestion
Update learned model
Interactive ML autonomous system
Heuristic based autonomous system
Expert observed user testing
Heuristics update
Heuristic-based autonomous system
in-the-wild user testing & evaluation
IML as participatory design
Baseline participatory design

LEARNT POLICIES



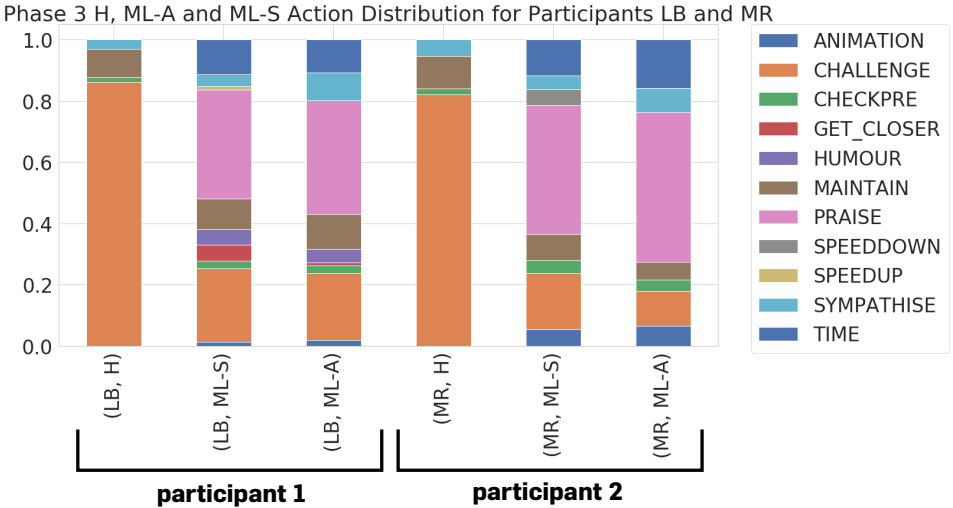
[Winkle et al. **In-Situ Learning from a Domain Expert for Real World Socially Assistive Robot Deployment** RSS 2020]

LEARNT POLICIES



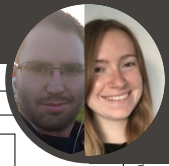
[Winkle et al. **In-Situ Learning from a Domain Expert for Real World Socially Assistive Robot Deployment** RSS 2020]

LEARNT POLICIES

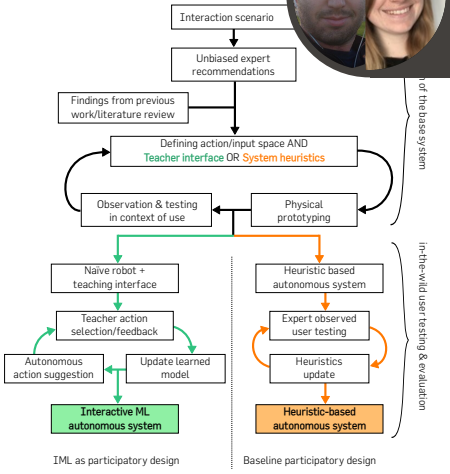


[Winkle et al. **In-Situ Learning from a Domain Expert for Real World Socially Assistive Robot Deployment** RSS 2020]

LEADOR: END-TO-END CO-DESIGN



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



[Senft et al. **SPARC: Supervised progressively autonomous robot competencies**, ICSR 2015;
Winkle et al. **LEADOR: End-to-End Participatory Design of Autonomous Social Robots** FrontiersIn AI & Robotics 2021]

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."

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."

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

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."

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

...is it *Responsible Robotics©*?

DOES IT TRANSLATE TO INDUSTRY?

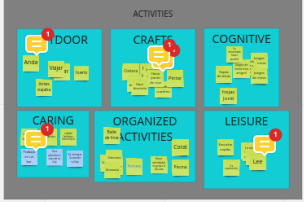
Can we turn the LEADOR paradigm into a generic & robust enough technique for broad usage?

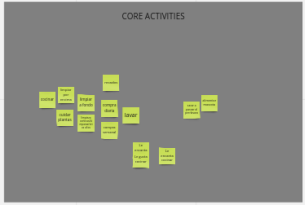
DOES IT TRANSLATE TO INDUSTRY?

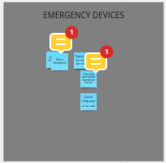
Can we turn the LEADOR paradigm into a generic & robust enough technique for broad usage?

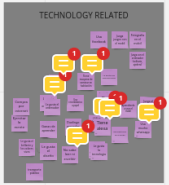
Two key use-cases for PAL Robotics:

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









NHoA: Never HOme Alone

Clínica Humana

Older person home (terrace)

Can Granada sheltered apartment

Older person home (reception)

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 exercice 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?

PAL ROBOTICS

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?

PAL ROBOTICS

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?

PAL ROBOTICS

`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)

Thank you!

We are always looking for great people to join us!

Drop me a line if you want to know more

PAL ROBOTICS