Exploring the Impact of Fault Justification in Human-Robot Trust

Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S. Melo, Ana Paiva



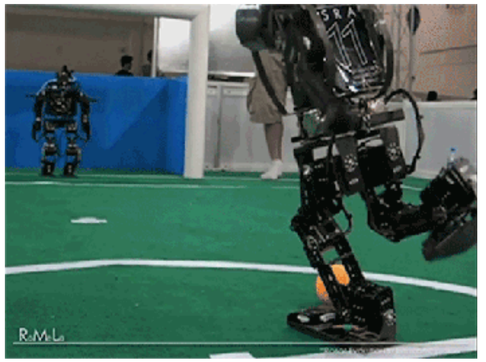


INESC-ID & Instituto Superior Técnico, Lisbon University, Portugal

Motivation

Motivation

- Robots fail even in controlled settings!



Motivation

- There are 2 types of error situations [1]:
  - Social norm violations
  - Technical failures

[1] Giuliani, M., Mirnig, N., Stollnberger, G., Stadler, S., Buchner, R. and Tscheligi, M., 2015. Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Frontiers in psychology*, *6*, p.931.

Motivation

- Error situations may:
  - Affect the perceptions/expectations of robots
  - Compromise the task
  - Cause frustration on the user
  - (…)

Motivation

Some faulty behaviours cannot be avoided…

…but can be detected!

What should a robot do after detecting a faulty behaviour?

Questions

- In **collaborative tasks**, how much do we **trust** a robot that had a **technical failure?**

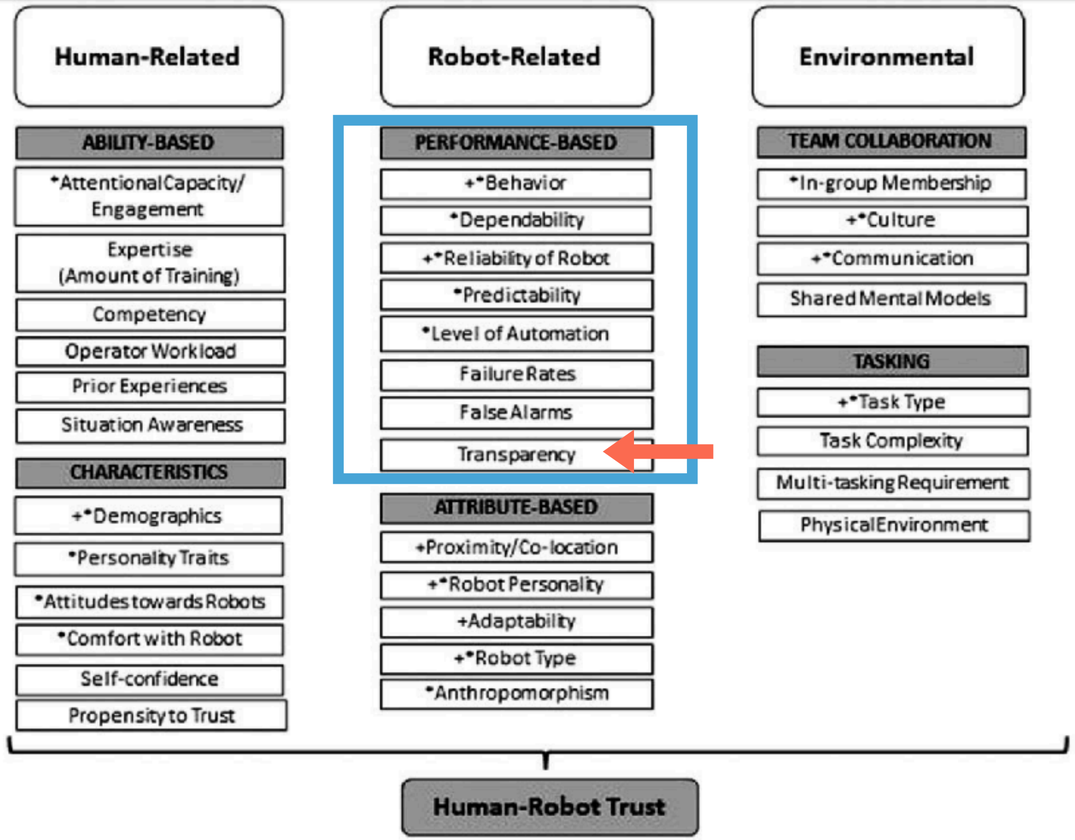
Questions

- In **collaborative tasks**, how much do we **trust** a robot that had a **technical failure?**

- What **recovery strategies** should a robot adopt in order to **mitigate** its negative effect?

Human-Robot Trust

"We define trust as the reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others" [1]



[1] Schaefer, K., 2013. The perception and measurement of human-robot trust. (Doctoral dissertation, University of Central Florida Orlando, Florida).

Questions

- In **collaborative tasks**, how much do we **trust** a robot that had a **technical failure?**

- Can the **justification recovery** strategy **mitigate** the negative effect of a technical failure?

Hypotheses

Hypotheses

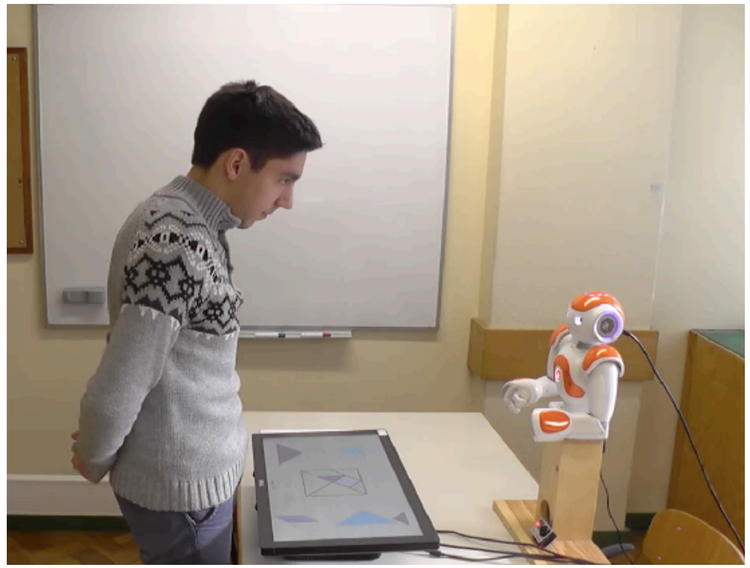
- **H1**: A **technical failure** of a social robot in a collaborative task will have a **negative effect on the trust** towards the robot.

- **H2**: A social robot that reveals transparency by **justifying the technical failure** during a collaborative task will **mitigate** the negative effect on the **trust** towards the robot.

User Study

Scenario

- Solve 3 tangram puzzles
- Solve collaboratively in turns
- Robot is autonomous
  - Simulates a technical failure
  - Simulates an autonomous recovery



Simulation of a Technical Failure

- The robot **stutters** “It’s myyyyyyyyyy” and **freezes** for 50 seconds.
    - We tuned the freezing time in a pilot experiment.

Independent variables

Recovery Strategy	“There was a failure in my speech module”
No recovery	**Justification**

Independent variables

Recovery Strategy / **Failure consequence**	No recovery	Justification
Task continues		
Task restarts		

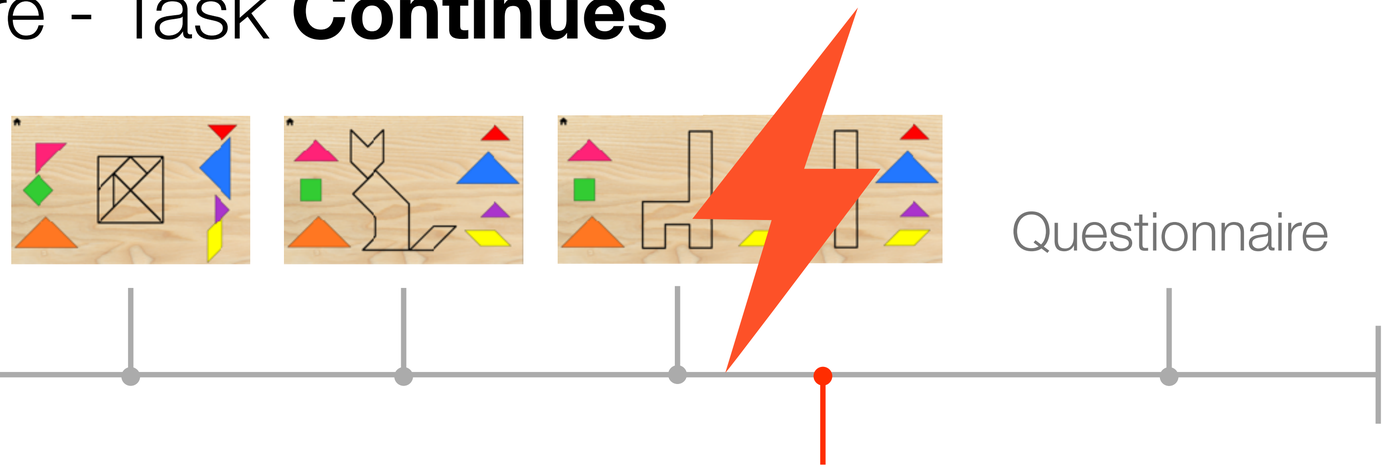
Independent variables

Recovery Strategy / Failure consequence	No recovery	Justification
Task continues	No recovery & Task continues	Justification & Task continues
Task restarts	No recovery & Task restarts	Justification & Task restarts

Experimental Design

- Between-subjects design
- 5 conditions
  - **Control (No failure!)**
  - Justification & Task Continues
  - Justification & Task Restarts
  - No Recovery & Task Continues
  - No Recovery & Task Restarts

Procedure - Task **Continues**



Procedure - Task **Continues**

Questionnaire

Questionnaire

Justification / No Recovery

Procedure - Task **Restarts**

Questionnaire

Justification / No Recovery

Questionnaire

Measures

- Human-Robot Trust Questionnaire (14-items sub-scale) [1]

- Impact of the failure on the task (manipulation check of the failure consequence):

  > “Identify the impact of the failure on the task from 1 (Not severe) to 5 (Very much severe)”

[1] Schaefer, K., 2013. The perception and measurement of human-robot trust. (Doctoral dissertation, University of Central Florida Orlando, Florida).

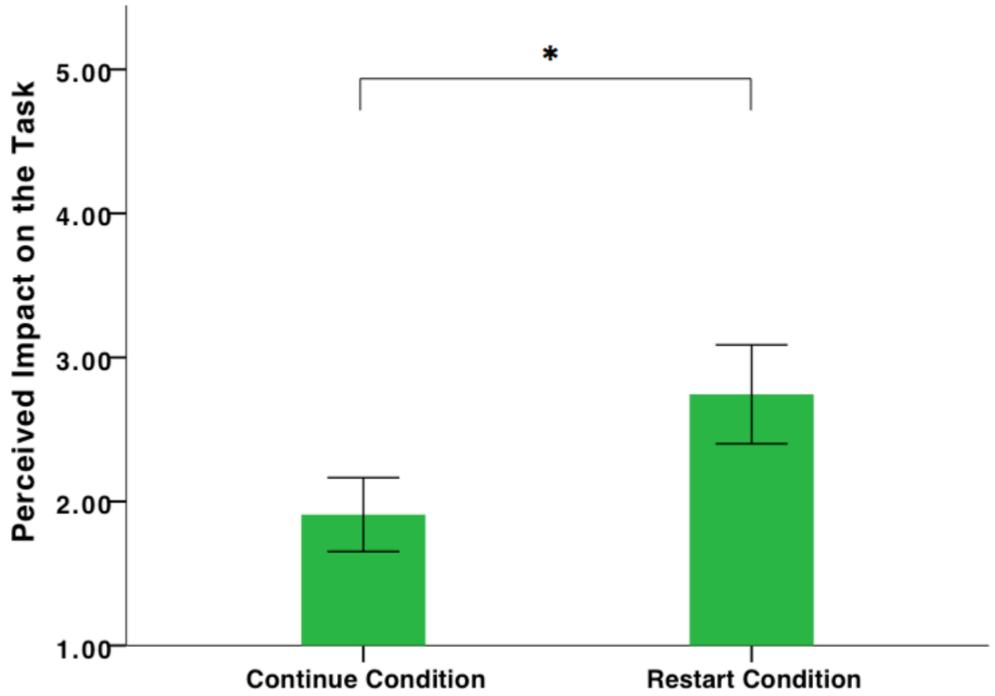
Participation

- 97 participants (71 males and 26 females, M_{age}=22.26±4.51)
  - 16 - Control
  - 18 - Justification & Task Continues
  - 21 - Justification & Task Restarts
  - 20 - No Recovery & Task Continues
  - 22 - No Recovery & Task Restarts

Results

Manipulation check of Failure Consequence

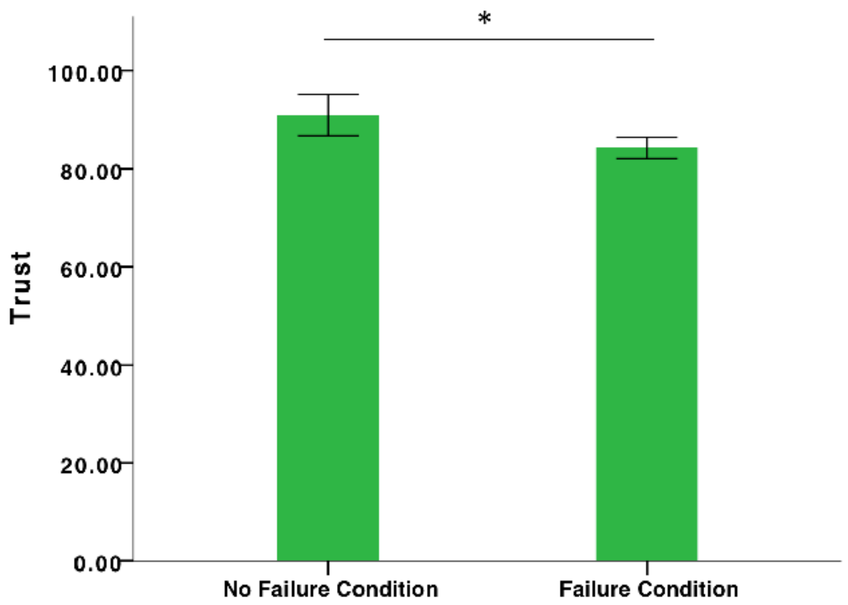
- Participants in the "**Task Continues**" group perceived the **failure as less severe** when compared to participants in the "Task Restarts" group.



Mann-Whitney Statistical Test
(U = 394, p = 0.001, r = 0.38)

Results - Trust

- Participants in the **Control group** showed **higher trust** levels towards the robot than the group where the robot presented the technical failure.



1-way ANOVA Statistical Test
(F = 12.97, p = 0.001)

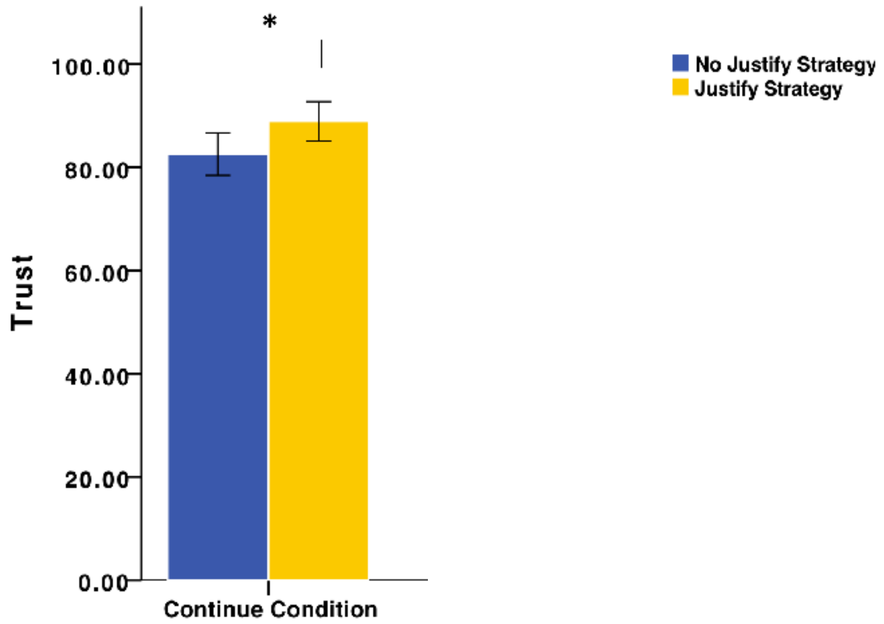
Verifies the homogeneity of variances assumption (Levene's test (p=0.998).

Results - Trust

- There was a significant **interaction effect** between the **Recovery Strategy** and the **Failure Consequence** [Factorial ANOVA Statistical Test (F = 4.17, p = 0.045)]

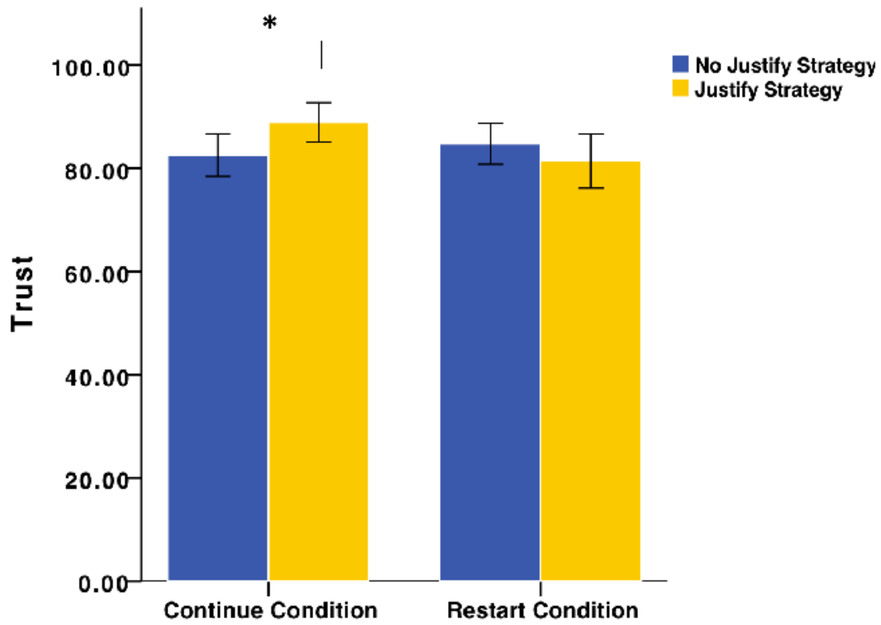
Results - Trust

- In the "**Task Continues**" group, participants reported **higher levels of trust towards** the robot that applied the **Justification** Strategy **than** towards the robot that applied **No Recovery** Strategy [Mann-Whitney Statistical Test (U = 107, p = 0.033, r = 0.35)].



Results - Trust

- In the "**Task Restarts**" group, there were **no statistically significant differences** between the trust levels towards the robot in both recovery strategies [Mann-Whitney Statistical Test (U = 201.5, p = 0.473)].



Results - Mitigation of Trust

The trust levels towards the robot were significantly different between the Control group and:

- No Recovery & Task Continues ($U = 77$, $p = 0.007$);
- Justification & Task Restarts ($U = 88.5$, $p = 0.013$);
- No Recovery & Task Restarts ($U = 108$, $p = 0.045$).

The trust levels towards the robot in the **Control group** and in **Justification & Task Continues** were not significantly different ($U = 119.5$, $p = 0.403$).

Discussion

- **H1**: A **technical failure** of a social robot in a collaborative task will have a **negative effect on the trust** towards the robot.

Discussion

H1: A **technical failure** of a social robot in a collaborative task will have a **negative effect on the trust** towards the robot.

Discussion

H1: A **technical failure** of a social robot in a collaborative task will have a **negative effect on the trust** towards the robot.

- **H2**: A social robot that reveals transparency by **justifying the technical failure** during a collaborative task will **mitigate** the negative effect on the **trust** towards the robot.

Discussion

H1: A **technical failure** of a social robot in a collaborative task will have a **negative effect on the trust** towards the robot.

H2: A social robot that reveals transparency by **justifying the technical failure** during a collaborative task will **mitigate** the negative effect on the **trust** towards the robot.

Conclusions

Conclusions

- We extended previous literature by analysing a **technical failure** and **unexplored recovery strategy** during a **collaborative task**.

- **Mitigation** strategies should be **tailored according to different factors** (e.g., task type, failure type, failure severity).

Conclusions

- **Technical failures** by a social robot in a collaborative task are perceived as **less trustworthy**.

- The **Justification** Strategy can **repair trust** in less severe failure consequences.

Some reactions to the recovery strategies

Thank you!

Questions?

filipacorreia@tecnico.ulisboa.pt