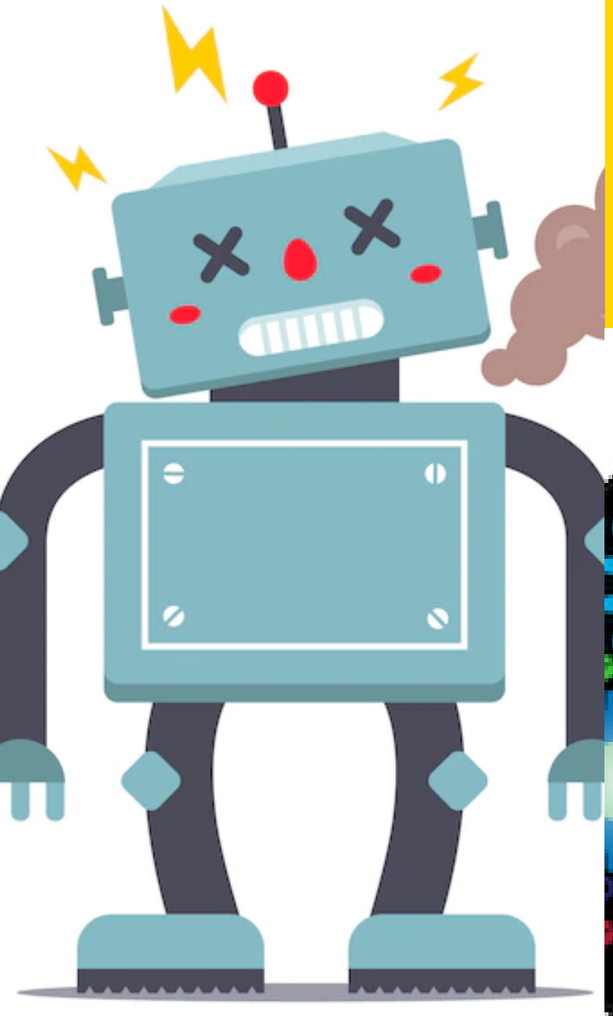
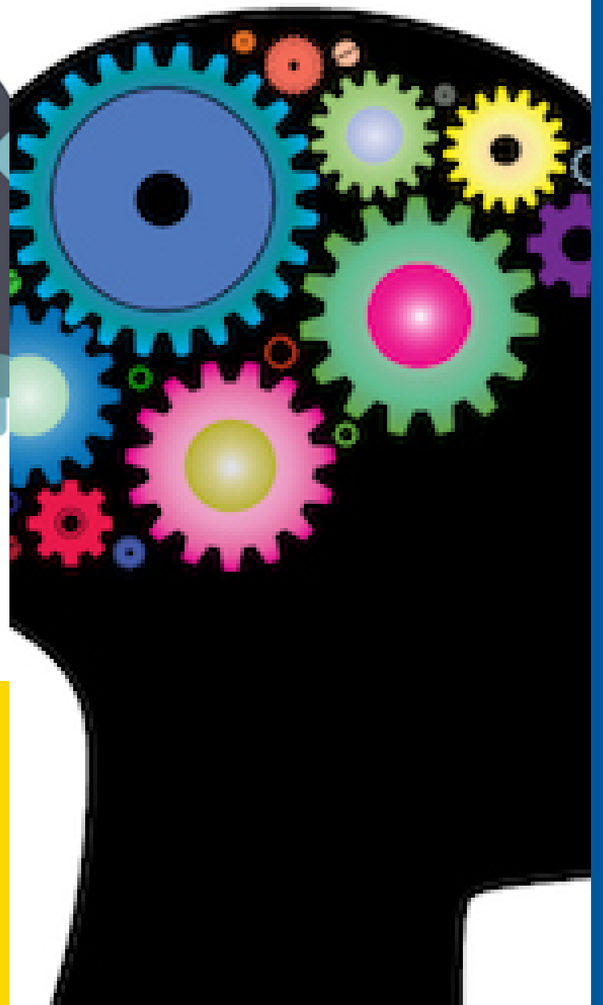


Defense
provided by
Lois Brun



0361

Evaluating the Effect of Theory of Mind on People's Trust in a Faulty Robot

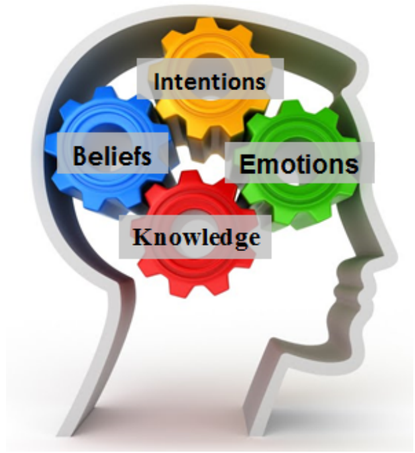
BRUN Loïs

INSTITUT POLYTECHNIQUE DE PARIS
TELECOM SudParis

Summary

Definition : ToM

Theory of Mind(ToM):



Multi-modal system that allows people to naturally communicate and understand each other by inferring others' intentions, desires, and beliefs

↓

Results in people's expectation of a similar ability when interacting with social robots.

Definition : Trust

Trust : essential in interpersonal and economic interactions, both in Human-Human (HHI) and Human-Robot (HRI) Interactions.

Trust : strongly related to the perception of reliability and to the people's willingness to take the risk to help the counterpart

Trust : can be destroyed by unexpected behaviors (slow navigation) or errors (mechanical malfunction)

Importance of ToM

ToM :

- allows people to establish the reliability of the information and capabilities of a robot, which helps to build people's trust in robots.

- allows robots to evaluate the situational context and plan accordingly to achieve a goal either in collaboration with humans or without any human direct intervention.

State of the art

- Few studies
- Looking into the effects of ToM on people's trust in robots
- State-of-art literature showed that several repair mechanisms exist that allow robots to recover people's trust after a breach

State of the art

- These studies also imply that the robot has awareness of itself and the task, but not necessarily that other agents have a ToM

- No previous studies investigated the relationship between robot errors in correlation to the robot's ability of displaying ToM skills

The importance of this study

- Robot recognises its **errors** and **communicates** its intention of **rectifying** the situation, it can recover from the negative effects of such errors on people's trust
- Other strategies also include **letting the robot explains the errors**, either by providing justifications for the failure or by providing a higher level of information
- Others also include the possibility of prompting the intervention of a human in support of the robot

The goal of this paper

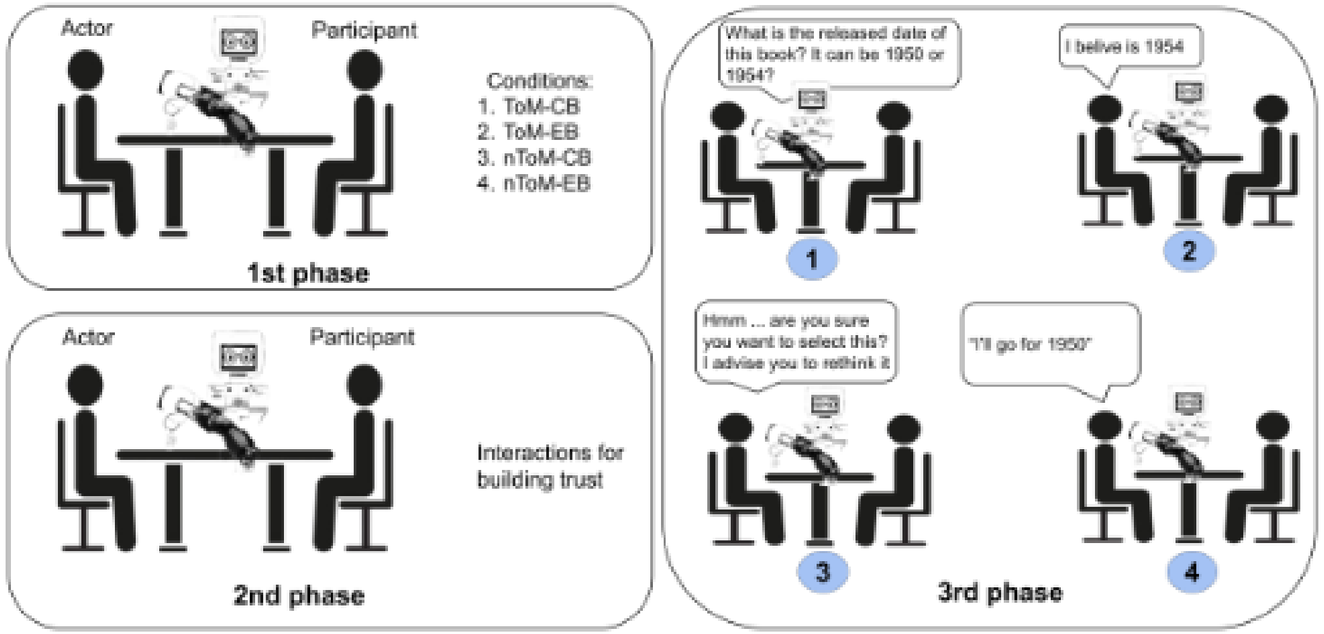
- Investigate whether the presence of ToM in a robot would affect people's trust in the robot even when the latter makes mistakes.

- Design a between-subjects study where participants played with a robot in a variation of a known guessing game

hypothesis : The presence of ToM in a robot may be a mitigating factor for a loss of people's trust in the robot, similar to a human-human trust recovery.

The experiment : variation of the Price Game

A robot engaged participants in a three-phase interaction, exhibiting behaviour that varied according to its ToM and erroneous (EB) or flawless behaviour (CB).



The experiment : variation of the Price Game

3 phases :

1	The robot was the player that was requested to guess the date with the suggestions provided by the human assistant
2	Robot engaged the participant and the actor in some small and flawless talks about books. In particular, this phase has been conceived to allow the robot to regain people's trust in its capabilities in the conditions with error
3	Aimed to assess people's trust in the robot. Specifically, the participant was requested to guess the correct date with a hint provided by the robot.

The experiment

experimental conditions:

1) ToM-CB : the robot had ToM and did not make any errors;

2) ToM-EB : the robot had ToM and made errors;

3) nToM-CB : the robot did not have ToM and had flawless behaviour;

4) nToM-EB : the robot had not ToM and had erroneous behaviour.

The experiment

1) ToM-CB:
the robot detected the actor's lie
("I believe that you thought to give me an incorrect answer about the date")
and guessed the correct answer
("The correct answer is 1987"),
2) ToM-EB:
the robot detected the lie and responded incorrectly
3) nToM-CB: the robot did not detect the lie and responded correctly by providing the correct answer
4) nToM-EB: the robot did not detect the lie and answered incorrectly ("The correct answer is 1989").

The Questions

Do you think that the robot was autonomous?
Y : 50,5%, N : 24.3%, U : 25,3%
Do you think that the robot already knew the answer?
Y : 30%, N : 48% U : 22%
Do you think that the robot lied to you?
Y : 27%, N : 46,5% U : 26,5%
Do you think that the robot detected the lie it was told?
Y : 41%, N : 30% U : 29%

Analysys of the results

Statistically significant effect of the robot's ToM on participants' belief that the **robot could understand that the actor lied to it**

No statistically significant association between the robot's flawless or erroneous behaviour and the questions

Condition with	Participants' Choice		
	Yes	No	I do not know
ToM	1.8	-2.5*	0.5
noToM	-1.8	2.5*	-0.5

THE ADJUSTED STANDARDISED RESIDUALS OF THE CROSS TABULATION BETWEEN THE PARTICIPANTS' BELIEF THAT THE ROBOT RECOGNISED CORRECTLY WHEN IT WAS LIED AND THE CONDITIONS OF ROBOT'S TOM. THE VALUES MARKED BY * ARE STATISTICALLY SIGNIFICANT (I.E., VALUES GREATER THAN 1.96 AND LOWER THAN -1.96).

First observations

Correlation between the robot having a ToM and a decrease in participants' belief that the robot did not recognise the lie.

The participants' choices were more affected when the robot had a theory of mind.

Participants correctly recognised that the robot had a ToM in the conditions where the robot's ToM was activated.

Trust measure

1) Observed trust in the robot

- (75.6%) accepted the robot's judgement and decided to change their answer
- 30% of the participants accepted to change their decision even if they guessed the correct answer to the game

2) Perceived reliability in the robot

- No statistically significant interaction between ToM and robot's behaviour on their perceived reliability of the robot
- A robot with ToM is generally considered more reliable than one without ToM even if it has erroneous behaviour

3) Percieved faith in the robot

- Tendency that participants have overall higher faith in a robot with ToM than in one without ToM even if it has erroneous behaviour

Trust measure

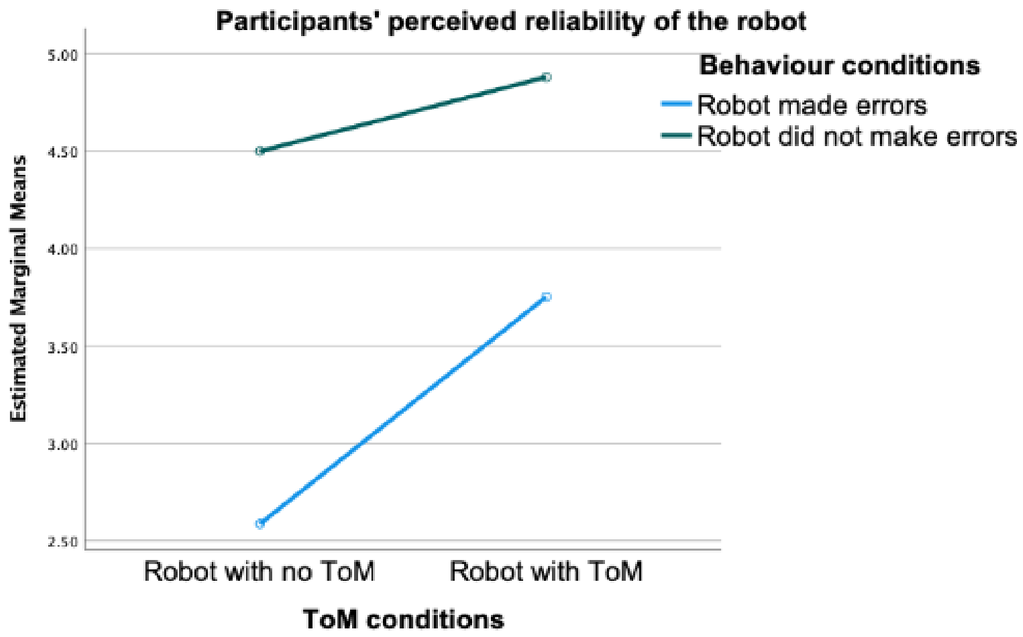


Fig. 4. Means of participants' perceived reliability of the robot.

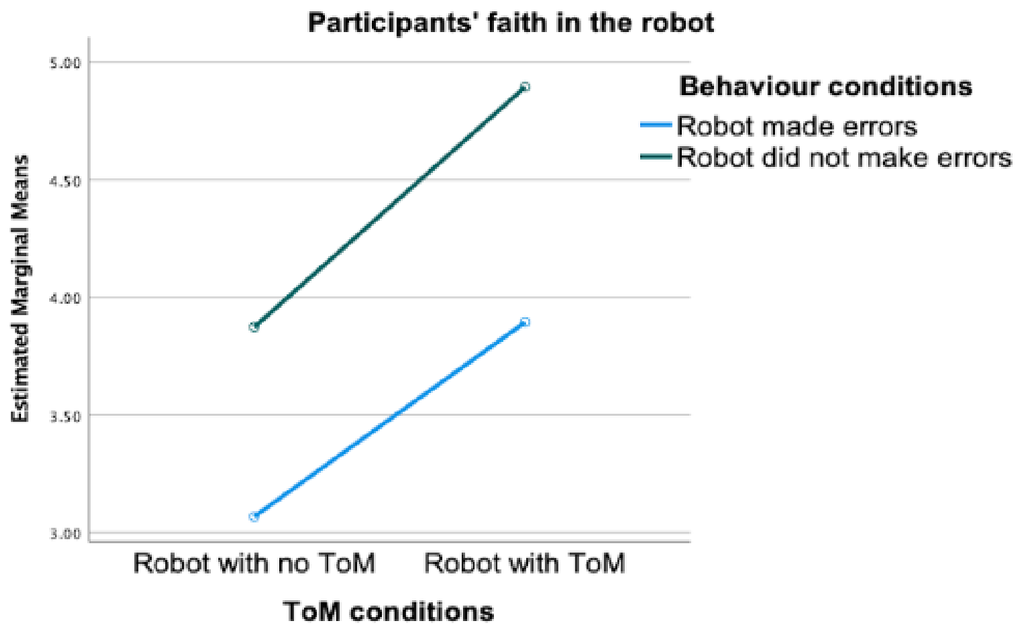


Fig. 5. Means of participants' faith in the robot.

Conclusions

We : investigated whether the ability of the robot to display ToM would affect humans' trust and their perception of the robot even when the latter does not always complete its tasks correctly.

We : designed a user study in a controlled but real-world setting in which we manipulated two variables: the robot's ToM and its behaviour

We : Showed that there is a positive correlation between the robot's ToM and participants' belief that the robot recognised the lie.

Summary & source

This study highlights the importance of ToM on people's perception of the robot by providing a powerful factor to manipulate their expectations and trust in the robot

Evaluating the Effect of Theory of Mind on People's Trust in a Faulty Robot

Alessandra Rossi[1]* Antonio Andriella[2]* Silvia Rossi[1] Carme Torras[3] Guillem Alenyà[3]