Investigating human perception of trust and social cues in robots for safe HRI in home environments

Alessandra Rossi[1] and Kerstin Dautenhahn[1,2] and Kheng Lee Koay[1] and Michael L. Walters[1]

Abstract— **Our aim is to create guidelines that allow humans to trust robots that are able to look after their well-being by adopting human-like behaviours. However, trust can change over time due to different factors, e.g. due to mechanical, programming or functional errors. It is therefore important for a domestic robot to have acceptable interactive behaviour when exhibiting and recovering from an error situation. As a first step, we investigated human users' perceptions of the severity of various categories of potential errors that are likely to be exhibited by a domestic robot. We conducted a questionnaire-based study, where participants rated 20 different scenarios in which a domestic robot made an error according to their severity. We clearly identified scenarios that were rated by participants as having limited consequences ('small' errors) and that were rated as having severe consequences ('big' errors). In order to define acceptable behaviours to recover the human trust, it is necessary to consider that errors can have different degrees of consequences and people's personalities and dispositions of trust may affect differently their perception of the robot. We used an interactive storyboard presenting ten different scenarios in which a robot performed different tasks, either correctly, or with small or big errors, under five different conditions. At the end of each experimental condition, participants were presented with an emergency scenario to evaluate their current trust in the robot. We conclude that there is correlation between the magnitude of an error performed by the robot and the corresponding loss of trust of the human in the robot. We also found a correlation both between individual personalities and characteristics of people and their perceptions of the robot and trust towards a robot.**

I. INTRODUCTION

In the not too distant future, autonomous robots will take part in peoples' daily living activities. In particular, humans will have to interact with them in domestic environments. This prospect will open two main challenges for consideration: Humans will need to accept the presence of the robot and they will also have to trust that their robotic companion will look after their well-being without compromising their safety. Trust determines human's acceptance of a robot as a companion and in their perception of the usefulness of imparted information and capabilities of a robot [1], [2]. Higher trust is associated with the perception of higher reliability [3]. Furthermore, other aspects such as the appearance, type, size, proximity, and behaviour of a particular robot will also affect user's perceptions of the robot [4], [5]. Syrdal et al. [6] showed that dog-inspired affective cues communicate a sense of affinity and relationship with humans. Martelaro et al. [7] established that trust, disclosure, and a sense of companionship are related to expressiveness and vulnerability. They showed how a sense of the robot's vulnerability, through facial expressions, colour and movements, increased perceived trust and companionship, and increased disclosure. Lohse et al. [8] demonstrated that robots with more extrovert personalities are perceived more positively by some users.

Robots are machines and they might exhibit occasional mechanical or functional errors. For example, the robot may turn off during a delicate task because its battery was fully discharged without warning, or a robot might unlock the front door to strangers who may be potential thieves. People might perceive errors differently according to the resultant consequences and the timing of when they happened. Indeed, the impact of 'big errors' or an accumulation of 'small errors' might be perceived differently.

Our works [9], [10], [11] analysed human users' perceptions of the severity of errors made by a robot and their impact on human users' trust. Such analysis was intended to categorise potential errors that are likely to be exhibited by a domestic robot according the participants' perceptions (i.e., which errors are considered having 'big' and 'small' consequences), and to identify how the timing and severity of these errors influence the participants' trust in robots. We analysed how human users' personalities and characteristics affect their trust towards robots. This is particularly relevant in designing guidelines for Human-Robot Interaction in home environments where the interaction is strictly connected to humans' dynamics.

Research Questions

This work has been carried out considering different assumptions to investigate the following research questions (R) and hypothesises (H):

R1 Which kind of erroneous behaviours impact a human's trust in a robot? **H1** We expect that there is a correlation between the magnitude of the error performed by the robot and the loss of trust of the human in the robot. We hypothesise that errors with severe consequences have more impact on humans' trust in robots.

R2 Does the impact on trust change if the error happens at the beginning or end of an interaction? **H2** We expect that there is a correlation between the timing in which the error is performed during the interaction and the loss of trust. Similar to Human-Human relationships [12], we believe that humans

[1]A. Rossi, K. Dautenhahn, K. L. Koay and M. L. Walters are with Adaptive Systems Research Group, School of Computer Science, University of Hertfordshire, Hatfield, UK [a.rossi, k.dautenhahn, k.l.koay, m.l.walters]@herts.ac.uk

[2]K. Dautenhahn is with the Departments of Electrical and Computer Engineering/Systems, Design Engineering, University of Waterloo, 200 University Ave. W. Waterloo, Ontario kerstin.dautenhahn@uwaterloo.ca

recover trust more completely and quickly after the violation of trust in a later stage of the Human-Robot relationship.

R3 Is it easier to recover/regain human trust when it is a big error that occurs either at the beginning or at the end of the interaction? Or is it easier to regain/recover it when a loss of trust is caused by a small error happening at either the ends of the interaction? **H3** We expect that there is a correlation between the time at which the error occurred and the magnitude of the error. We hypothesise that a big error has more impact on the loss of trust when it happens at the end of the interaction because the human users do not have time to recover from the loss of trust.

R4 Do personalities and characteristics of humans affect their perception of a robot? Do personalities and characteristics of humans affect their trust in a robot? **H4** We expect that there is a correlation between both the personalities and characteristics of people, their perception of the robot and their trust in a robot. As with Human-Human relationships [13], [14], [15], we hypothesise that people with stronger and more positive attitudes towards other humans are more likely to trust robots.

R5 Are the use of human social behaviours sufficient for humans to trust a robot to look after their well-being? **H5** We believe that social cues make robots more human-like, and better accepted by humans, then humans can be more inclined to rely on them.

R6 Can a human's trust in her robot change over time? **H6** We believe that trust could change if the initial conditions of trusting a robot change, e.g. the robot starts to show erratic behaviours.

II. Human perceptions of the severity of domestic robot errors

There are several definitions of trust, however there is a tendency [17] in adopting the following definition: "Trust can be defined as the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability" [18, p. 51]. Trust is a complex feeling even between humans [16] and it can change during the course of interactions due to several factors [1].

Higher trust is associated with the perception of higher reliability [3]. Therefore, humans may perceive erroneous robot behaviours according to their expectations of a robot's proper functions [19]. However, robots can be faulty, due to mechanical or functional errors. For example, a robot might be too slow due to batteries running low. It might not be able to detect an obstacle and destroy a human user's favourite object, or the arm of the robot might cause a breakage during a delicate task. Each of these examples are robot errors, though their magnitude might be perceived differently according to the resultant consequences.

But which type of errors have more impact on human perceptions of robots? Factors may include severity and duration, the impact of isolated 'big errors', or an accumulation of 'small errors'. For example, Muir and Moray [31] argue that human perceptions of a machine are affected in a more severe and long-term way by an accumulation of 'small' errors rather than one single 'big' error. The embodiment of a robot may also have a major impact on the perception of it by humans [4].

What is perceived as a 'big error' and what is a 'small error'? People have individual differences, including age, gender, cultural and social habits, which may impact their perceptions of what are considered big or small errors. In order to study the differences in terms of the impact of errors on a human-robot interaction, first we have to establish what people consider subjectively to be 'small' or 'big' errors exhibited by a home companion robot. In this context, our first study was directed towards the classification of likely robot errors according to their perceived magnitude.

A. Method

This study has been organised as a within-subjects experiment. Each participant has been shown the same questions, rated using a 7-point Likert scale [1= small error and 7=big error].

B. Procedure

Participants were asked to imagine that they live with a robot companion in their home. However, the robot might make some mistakes. The participant has to complete a questionnaire rating the magnitude of the errors illustrated in different scenarios, e.g. "Your robot leaves your pet hamster outside the house in very cold weather". The questionnaire is composed of 20 questions, plus two optional in which the participant is free to add their own examples of errors not already included in the scenarios proposed.

C. Results

According to the resulting answers of 50 participants - (32 men, 18 women), 19 to 63 years old [mean 41, std 11.59]. All the questions with values < 4 are considered small errors, those with values > 4 are considered big errors and those with values $= 4$ are considered neutral errors. We identified 7 big errors, 6 small errors and 7 moderate errors. We did not find any significant differences between gender or age of the participants and their rating of the errors.

III. How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario

In order to enable safe Human-Robot Interaction in home environments, it is important to investigate how an interactive relationship can be established and preserved between human users and their robotic companions, along with the likelihood of robot errors occurring. In this context, this study investigated the impact of errors with different magnitudes and order of presentation on peoples' trust of robots.

A. Method

As part of a virtual, interactive storyboard, we observed and analysed participants' behaviours during interactions with a robot called Jace. We used a between-subject experimental design. Participants were asked to read a story and interact with the robot, using their mouse and keyboard,

whenever they were invited by the robot. In order to test our research questions, each experiment was executed under 5 different conditions: condition **C1**: 10 different tasks executed correctly by the robot; condition **C2**: 10 different tasks with 3 trivial errors at the beginning and at the end of the interaction; **C3**: 10 different tasks with 3 trivial errors at the beginning and 3 severe errors at the end of the interaction; **C4**: 10 different tasks with 3 severe errors at the beginning and 3 trivial errors at end of the interaction; and **C5**: 10 different tasks with 3 severe errors at the beginning and at the end of the interaction. All the conditions with errors were interspersed by the same 4 correct behaviours.

At the end of each condition, the participants were presented with a final task in which a fire started in their kitchen and they were presented with the following options 1) to trust the robot choosing the option "I trust Jace to deal with it."; 2) to not trust the robot choosing the option "I do not trust Jace. I will deal with it."; 3) to work with the robot, supervising the emergency, choosing the option "I want to extinguish it together with Jace."; 4) to not trust either the robot or themselves choosing the option "We will both leave and call the fire brigade.".

Finally, in order to analyse the interaction between the human participants and the robot, we asked the participants to answer two sets of different questions.

B. Procedure

Participants were asked to imagine that they lived with a robot as a companion in their home which helps them with everyday activities. They were tested using an interactive storyboard accessible through a web application.

We asked participants different questions at the beginning and end of the interaction:

Questionnaire 1 A pre-experimental questionnaire for 1) collecting demographic data (age, gender and country of residence), 2) the Ten Item Personality Inventory questionnaire about themselves (TIPI) [20], 3) 12 questions to rate their disposition to trust other humans [21] and 4) and to assess participants' experience and opinion with regard to robots.

Questionnaire 2 A post-experimental questionnaire including: 1) questions to confirm that participants were truly involved in the interactions and had noticed the robot's errors, 2) to collect participants' considerations about their feelings in terms of trust and appeasement (e.g."was the robot irritating/odd?" and "why did/did not you trust the robot?"), and their perceptions of the interactions (e.g. "did the scenario look realistic?") and 3) questions to collect the participants' evaluation of the magnitude of the errors presented during the interactions.

C. Results

We analysed responses from 200 participants (115 men, 85 women), aged 18 to 65 years old [avg. age 33.56, std. dev. 9.67]. Participants' country of residence was: 60% USA; 34% India; 6% European and other countries.

We asked participants four questions about the content of the scenarios to verify the level of their engagement with the story presented. Correct answers were received for 79.75% (max 92%, min. 71.5%). We analysed the responses of 154 participants, not including those who gave more than one wrong answer (thus identified as not paying very much attention to the study - which can be expected in an online survey) to the verification questions.

We observed that a majority of participants chose to deal with the emergency situation collaboratively, and a slightly smaller majority chose to trust the robot when tested with **C1**. Participants chose not to trust the robot when it made severe errors (**C5**), while they were more inclined to trust in teamwork when the robot made small errors (**C2** and **C3**). We also noticed that the number of participants who chose to trust the robot increased in **C3**. While this might indicate a tendency of participants to not trust the robot more when the severe errors were made by the robot at the beginning of the interaction, we did not find any statistically significant association.

We observed that the association of the choices of the participants for the emergency scenario and the experimental conditions is statistically significant ($\chi^2(12) = 32.91, p = 0.001$). The strength of relationship (Cramer's V) between the emergency choice and experimental conditions is moderate ($\phi_c = 0.26, p = 0.001$).

There is a correlation between the condition **C5** and the choice of the participants to not trust the robot (adjusted value > 1.96). We observed that participants' trust is affected more severely when the robot made errors with severe consequences. We did not find any significant dependency ($p > 0.3$) between the gender of the participants and their choices in trusting the robot to deal with the emergency. We did not find any statistically significant association for different age ranges of the participants and their emergency choices ($p > 0.12$). Therefore, we assume that these results can be generalised to a generic population independently of gender and age. Moreover, in order to test the association between participants' emergency choices and their country of residence, we used a Chi-Square Test. Since the majority of the countries of residence had only one participant, we applied the test only to India and USA. We observed that the association is not statistically significant ($\chi^2(3) = 4.138, p > 0.24$)).

We found a strong connection between the personality traits of agreeableness, conscientiousness and emotional stability, and their disposition of trust other people.

The majority of our participants did not have any previous experience of interaction with robots (79.97%, min=1, max=6, mean 1.64, std. dev. 1.27). Interestingly, from participants' responses we noticed that according to their experiences, extroverted participants tended to consider robots generally as a machine ($p = 0.007$) and agreeable participants as an assistant ($p = 0.007$), in contrast to their perceptions of the robot they interacted with in this study. In particular, extroverts perceived Jace as a friend ($p = 0.0019$) and a warm and attentive entity ($p = 0.0025$), while agreeable participants perceived Jace as a tool ($p = 0.0033$). We also found that extroverted participants would like to have Jace

as home companion ($p = 0.001, r = 0.269$) and believe it is reliable ($p = 0.002, F = 2.729$) and trustworthy in uncertain and unusual situations ($p(12) = 0.026, F = 2.025$).

Finally, we analysed participants' personalities and dispositions of trust with regard to their final choice of trusting the robot in an emergency scenario. We found that conscientiousness ($p(3) = 0.42, F = 2.803$) and agreeableness ($p(3) = 0.022, F = 3.320$) traits correlate with participants' propensity for trusting the robot, and participants' belief in benevolence of people also correlate with higher trust in Jace ($p = 0.014, F = 6.078$). Moreover, we observed that the errors made by the robot significantly affected participants' perception of the robot.

IV. CONCLUSIONS

Regarding the research question **R1**, our hypothesis **H1** suggested that there is a correlation between the severity of the error performed by the robot and humans not trusting the robot. Our study shows that the magnitude of the errors made by the robot, and humans not trusting the robot are correlated. In particular, participants' trust was affected more severely when the robot made errors having severe consequences. We also hypothesised in **H2** that the timing when the error is performed affects the trust towards robots (research question **R2**), and there is a correlation between the timing of when the error occurred and the magnitude of the error (research question **R3** and hypothesis **H3**). Our results marginally suggest also that there might be a tendency not to trust the robot when severe errors happen at the beginning of an interaction, but these differences were not statistically significant.

As indicated in Hypothesis **H4**, we found a correlation both between individual personalities and characteristics of people and their perception of the robot and trust towards a robot (research question **R4**).

We are currently investigating research questions **R5** and **R6**.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642667 (Safety Enables Cooperation in Uncertain Robotic Environments - SECURE).

REFERENCES

[1] D. Cameron, J. M. Aitken, E. C. Collins, L. Boorman, A. Chua, S. Fernando, O. McAree, U. Martinez-Hernandez, and J. Law, "Framing factors: The importance of context and the individual in understanding trust in human-robot interaction," in *International Conference on Intelligent Robots and Systems*, 2015.

[2] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors: The Journal of Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, 2011.

[3] J. M. Ross, "Moderators of trust and reliance across multiple decision aids (doctoral dissertation), university of central florida, orlando." 2008.

[4] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The benefits of interactions with physically present robots over video–displayed agents," *International Journal of Social Robotics*, vol. 3, no. 1, pp. 41–52, 2011.

[5] K. L. Koay, D. S. Syrdal, M. L. Walters, and K. Dautenhahn, "Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 2007, pp. 564–569.

[6] D. S. Syrdal, K. L. Koay, M. Gcsi, M. L. Walters, and K. Dautenhahn, "Video prototyping of dog-inspired non-verbal affective communication for an appearance constrained robot," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 2010, pp. 632–637.

[7] N. Martelaro, V. C. Nneji, W. Ju, and P. Hinds, "Tell me more designing hri to encourage more trust, disclosure, and companionship," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2016, pp. 181–188.

[8] M. Lohse, M. Hanheide, B. Wrede, M. L. Walters, K. L. Koay, D. S. Syrdal, A. Green, H. Hüttenrauch, K. Dautenhahn, G. Sagerer, and K. Severinson-Eklundh, "Evaluating extrovert and introvert behaviour of a domestic robot -a video study," in *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, 2008, pp. 488–493.

[9] A. Rossi, K. Dautenhahn, K. L. Koay, and M. L. Walters, "Human perceptions of the severity of domestic robot errors," in *Social Robotics*, A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, and H. He, Eds. Cham: Springer International Publishing, 2017, pp. 647–656.

[10] ——, "How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario," in *Social Robotics*, A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, and H. He, Eds. Cham: Springer International Publishing, 2017, pp. 42–52.

[11] R. Alessandra, D. Kerstin, K. Kheng Lee, and M. L. Walters, "The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario." vol. 9, 2018.

[12] O. Schilke, M. Reimann, and K. S. Cook, "Effect of relationship experience on trust recovery following a breach," *Proceedings of the National Academy of Sciences*, vol. 110, no. 38, pp. 15 236–15 241, 2013.

[13] M. P. Haselhuhn, M. E. Schweitzer, and A. M. Wood, "How implicit beliefs influence trust recovery," *Psychological Science*, vol. 5, pp. 645–648, 2010.

[14] T. Mooradian, B. Renzl, and K. Matzler, "Who trusts? personality, trust and knowledge sharing," *Management Learning*, vol. 37, no. 4, pp. 523–540, 2006.

[15] F. B. Tan and P. Sutherland, "Online consumer trust: A multi-dimensional model." vol. 2, 2004, pp. 40–58.

[16] R. M. Kramer and P. J. Carnevale, "Trust and intergroup negotiation," *Blackwell Handbook of Social Psychology: Intergroup Processes (eds R. Brown and S. L. Gaertner)*, pp. 431–450, 2003.

[17] K. Yu, S. Berkovsky, R. Taib, D. Conway, J. Zhou, and F. Chen, "User trust dynamics: An investigation driven by differences in system performance," vol. 126745. ACM, 2017, pp. 307–317.

[18] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, 2004.

[19] M. L. Walters, M. A. Oskoei, D. S. Syrdal, and K. Dautenhahn, "A long-term human-robot proxemic study," in *2011 RO-MAN*, July 2011, pp. 137–142.

[20] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr., "A very brief measure of the big five personality domains. journal of research in personality," pp. 504–528, 2003.

[21] D. H. McKnight, V. Choudhury, and C. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," *Information Systems Research*, vol. 13, no. 3, pp. 334–359, 2001.