How the Timing and Magnitude of Robot Errors Influence Peoples' Trust of Robots in an Emergency Scenario

Alessandra Rossi(✉), Kerstin Dautenhahn, Kheng Lee Koay,
and Michael L. Walters

Adaptive Systems Research Group, University of Hertfordshire,
College Lane, Hatfield, UK
{a.rossi,k.dautenhahn,k.l.koay,m.l.walters}@herts.ac.uk

Abstract. Trust is a key factor in human users' acceptance of robots in a home or human oriented environment. Humans should be able to trust that they can safely interact with their robot. Robots will sometimes make errors, due to mechanical or functional failures. It is therefore important that a domestic robot should have acceptable interactive behaviours when exhibiting and recovering from an error situation. In order to define these behaviours, it is firstly necessary to consider that errors can have different degrees of consequences. We hypothesise that the severity of the consequences and the timing of a robot's different types of erroneous behaviours during an interaction may have different impacts on users' attitudes towards a domestic robot. In this study we used an interactive storyboard presenting ten different scenarios in which a robot performed different tasks under five different conditions. Each condition included the ten different tasks performed by the robot, either correctly, or with small or big errors. The conditions with errors were complemented with four correct behaviours. 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.

Keywords: Human-Robot Interaction · Social robotics · Robot companion · Trust in robots · Trust recovery

1 Introduction

In the future, autonomous robots will be ubiquitous both in humans' working and private environments. For example as assistants in hospitals, scouts in military scenarios and as home companions. In particular, in home environments these robots can improve people's safety by providing them with both cognitive and physical assistance. For example, a robot could warn its human companion about a fire started in the kitchen, or remind its older human companion to take

A. Kheddar et al. (Eds.): ICSR 2017, LNAI 10652, pp. 42–52, 2017.
https://doi.org/10.1007/978-3-319-70022-9_5

her daily pills, or it could warn its human companion that there is a broken glass on the floor that the toddler may walk on it. On the other hand, human users also need to trust that their robot is able to look after their well-being without compromising their safety. For example, humans should be able to wake up in the middle of the night to drink water, to switch on the light and rely that they will not stumble over their robot and be injured. Trust has a key role in any human's acceptance of a robot as a companion. Indeed, trust also affects the humans' perception of the usefulness of the information and capabilities of a robot [6,12]. Higher trust is associated with the perception of higher reliability [22]. 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 [3,14]. Robots may be faulty, due to mechanical or functional errors. For example, a robot might be too slow, the arm of the robot might cause a breakage during a delicate task, or a robot might talk about personal and relevant information with a stranger to its human companion without even being aware of it. Each of these examples are errors, though their magnitude might be perceived differently according to the resultant consequences. But which type of error has more impact on a human's perception of the robots? Factors may include severity and duration, the impact of 'big errors', or an accumulation of 'small errors'. For example, Muir and Moray [19] argue that humans' perception of the machine is affected in a more severe and long-term way by an accumulation of small errors rather than one single big error. However, the embodiment of a robot may have a major impact on the perception of it by humans [3]. Trust of people is not only affected by magnitude of the error made by the culprit but also by the type and length of the relationship they are in. "Will people be more likely to forgive a breach of trust in an earlier or later stage of an interpersonal relationships?" [26, p. 15236] Schilke et al. [26] investigated how certain kinds of relationships recover better and faster after a violation of trust, and how the timing of the violation affects the recovery. They demonstrated that people in longer relationships re-establish their mutual trust easier than those in newer relationships. Therefore with regard to robots we believe that the order of presentation of errors happening may affect differently the trust of human users in their robot companions. In this study we investigated the effect of order of presentation of robot errors on human trust towards robots during a Human-Robot Interaction.

2 Background and Related Work

Several studies define the concept of trust in Human-Human, in Human-Computer and Human-Robot Interactions. It is not clear which kind of errors, with trivial or severe consequences, have more impact on human users' trust towards robots [7,19,21,25]. Individuals also react very differently after a trust violation [13,26,27]. Some are quick to forgive while others believe that once the trust is broken the culprit cannot gain it back [29]. Therefore, this paper investigate how human users' overall trust in the robots is affected by robot's errors.

2.1 Definition of Trust

Humans rely on other humans everyday for different things. For example, students may trust their mentors to provide them with reliable information, and travellers trust the pilot flying an aeroplane to arrive at their destination safely. However, trust is a complex feeling [15] and it may also be influenced by different internal and external factors. For example, Simpson [27,28] highlights four core principles that affect trust: individuals assess the degree of trust by observing a partner acting unselfishly and supporting the best interests of both the individuals and the relationship. The individuals may purposefully create situations to test their partner's trust, and individuals with lower self-esteem may be less trustful of their partner. The level of trust in short-term or long-term relationships cannot be fully understood without considering the predisposition of trust of all the parties involved in the relationship. Mayer et al. [17] established that trust is constructed from a perception of ability, benevolence and integrity. Human-human trust is also affected by the perception of the risk of an interaction with other humans. The popular poker game is a concrete example to how risk-taking behaviours affect the credibility of a poker player, and thus it is important for all players to develop a good reputation during a game [4]. Deutsch [9] claims that risk-taking and trusting behaviour are different sides of the same coin, and that a person is willing to take a risk only if the odds of a possible positive outcome are greater than those for a potential loss. Golder and Donath [10] claim that a good reputation is very important in enhancing trust both in short and long term relations. Although multiple definitions exist, and several previous studies have adopted one of the first definitions of trust [9], there is a convergent tendency [31] towards using the definition "Trust can be defined as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [16, p. 51]. Lee [16]'s definition encapsulates the most important factors that can affect Human-Robot Trust: Human-related, Robot-related and Context-related.

2.2 Trust and Errors in Human-Robot Interactions

Bainbridge et al. [3] found that participants were happy to follow a robot's instructions to throw books in the trash if the robot was present in the room with them, but not when the robot was not physically in the same room. Other studies ([7,8]) showed that the order of presentation of the decreased reliability produces an evident drop in the trust in the robot which can be restored by continuing the interaction. They showed also that warning the participants about a drop in the robot's performance can mitigate the loss in trust. However, while in these studies the errors made by the robot have the same impact in terms of cost in the interaction, we argue there could be a different outcome according the severity of the error. Wang et al. [30]'s studies showed that the frequency and significance of errors can impact humans' trust in an imperfect on-line system. They showed that people are not willing to follow an imperfect robot if the outcomes are severe. No matter how close a human can feel to their avatar during an on-line interaction, the serious consequences of their actions do not have a great significance in real life. Booth et al. [5] investigated participants' responses to a robot's request to move in a secure-access student dormitory. They conducted the experiment with two conditions: (1) an anonymous robot and (2) a food delivery robot, where both asked to enter the building. They observed that participants were more likely to let the food delivery robot enter the building or in situations when they were in a group. Robinette et al. [20] investigated the effects of apologies, promises and additional reasons given by a robot for its errors on participants' trust in a simulated evacuation scenario conducted in a virtual environment. They showed that participants' trust was repaired if the robot apologised and promised to not repeat the error soon after it made the error but not during the emergency. Salem et al. [25] studied human perception of trust in robots, and how willing they are to follow a robot showing faulty behaviours. They showed that no matter how erratic the behaviour of the robots, participants followed the instructions of the robots. Similarly, Robinette et al. [21] used an emergency evacuation scenario, with artificial smoke and a smoke detector, in which a robot guided a person to an exit, in order to study how willing were humans to follow a robot that had previously exhibit erratic behaviour. Their results indicated that all the participants of the experiment followed the robot's instruction. In both experiments participants trusted the robots for different reasons. For example, some of them believed it was all staged, others that they were supposed to follow it because they accepted to participate in the experiments. Both Salem et al. [25] and Robinette et al. [21]'s works showed that some participants did believe that they were acting according to the experimenter decisions and that their lives where not in danger. Therefore, it is still not clear from these results whether faulty robots are trusted by humans, and whether humans can believe that robots can look after their safety and well-being.

3 Methodology

Investigating trust and human perceptions of safety in Human-Robot Interaction is not a simple task due to ethical concerns and risks [24]. Moreover, to fully investigate the impact of errors with different magnitudes it is important to create an interaction scenario with a fully-functional and versatile robot. For example, the robot should be able to grasp objects, to move autonomously, to detect obstacles and objects at runtime, to talk and to perform speech recognition. For an initial experiment, therefore, we decided to use an interactive storyboard through which participants interacted with a home companion robot called Jace.

3.1 Experimental Design

We used a graphical interface to observe and analyse participants' behaviours during the interaction with the robot. We used a between-subject experimental design. The participants were asked to read the story and interact, using their mouse and keyboard, whenever they were invited to by the robot. In order to test our research questions, each experiment was executed under 5 different conditions, as illustrated in Fig. 1: condition **C1**: 10 different tasks executed correctly; condition **C2**: sing 10 different tasks with 3 severe errors at the beginning and at the end of the interaction; condition **C3**: 10 different tasks with 3 severe errors at the beginning and 3 trivial errors at end of the interaction; condition **C4**: 10 different tasks with 3 trivial errors at the beginning and 3 severe errors at the ends of the interaction; and condition **C5**: 10 different tasks with 3 trivial errors at the beginning and at the end of the interaction. All the conditions with errors are interspersed by the same 4 correct behaviours. The classification of the robot's errors according to their magnitude has been validated in our precedent study [23], in which we asked participants to rate several errors made by a robot according their magnitude. An example of trivial error is 'You ask for a cup of coffee. Your robot brings you an orange.'. A severe error example is 'Your robot leaves your pet hamster outside the house in very cold weather.'. At the end of each condition, the participants were presented with a final task in which a huge fire started in their kitchen. In order to analyse the interaction between the human participants and the robot, we asked the participants different questions.

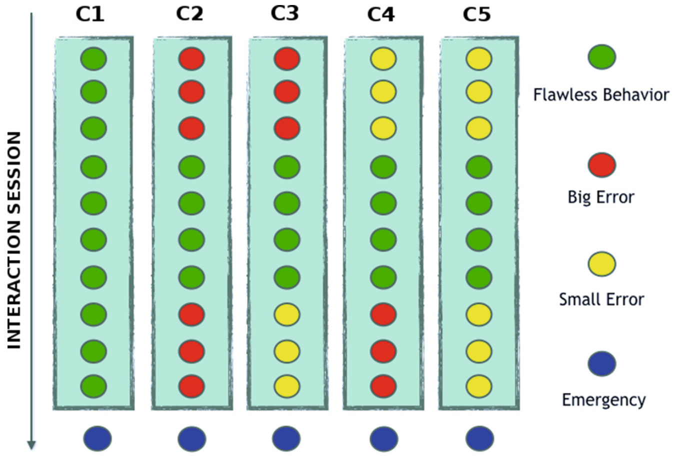


Fig. 1. Experimental conditions presented to the participants.

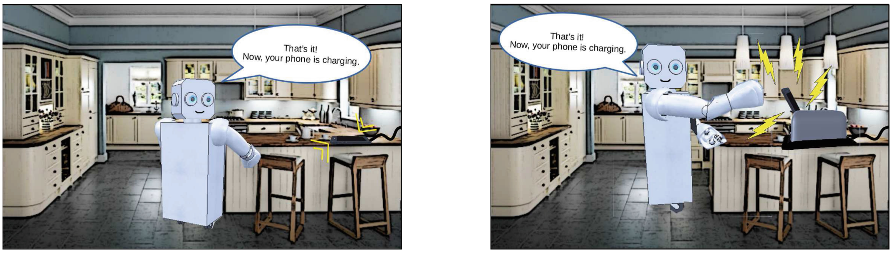
Questionnaire 1. A pre-experimental questionnaire for (1) collecting demographic data (age, gender and country of residence), (2) the Ten Item Personality Inventory questionnaire about themselves (TIPI) [11] and (3) 12 questions to rate their disposition to trust other humans [18], (4) and to assess 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?"), (3) questions to collect the participants' evaluation of the magnitude of the errors presented during the interactions.

Finally, objective measures were considered to confirm whether or not participants followed the robot's suggestions, i.e. observing the choices made during the emergency scenario.

3.2 Experimental 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. They were presented with 10 different scenarios, in which the robot showed flawless and erroneous behaviours. We chose the scenarios according to the results of our previous study [23]. Figure 2 shows an example of scenario in which the robot executes the required task (Fig. 2(a)) correctly putting the user's phone on charge and (Fig. 2(b)) putting the user's phone inside the toaster. At the end of each scenario, the participants were presented with an emergency situation, i.e. 'a fire in the kitchen' to finally assess their level of trust in the robot.



(a) *The robot puts the phone on charge.*

(b) *The robot puts the phone in the toaster.*

Fig. 2. The participant asks the robot to charge her phone. In the Figure (a) the robot does the task correctly; in the figure (b) the robot puts the phone in the toaster making a very dangerous error. In both figures, the robot believes that it had correctly performed the task and states this to the participant.

3.3 Participants

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; 1.5% Venezuela; 1.5% Portugal; 0.5% UK; 0.5% Canada; 0.5% Germany; 0.5% Dominican Republic; 0.5% Sweden; 0.5% Nigeria. The recruitment was carried out by using the crowd sourcing webservice Amazon Mechanical Turk [1]. These services are not used to replace live Human-Robot Interactions, but provide useful data in the early phases of a research project.

4 Results

We asked participants to rate their perception of the interaction. A seven point rating scale, ranged from 1 to 7 (disagree to agree), was used to measure the participants' judgement of the realism of the scenarios. Sixty-five percent of participants rated the scenarios as very realistic (rating values >4), 20% rated the scenarios as not realistic (rating values <4) and 15% neither agreed nor disagreed. We also 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%). However, for the question "Which secret did your robot Jace tell you?", 13% of the participants answered with the secret that they themselves had told the robot. We hypothesize that they misunderstood the question. 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.

All participants were presented with the same final emergency scenario. The options were been carefully chosen as indicators that the participant respectively trusts the robot, does not trust the robot, trusts in collaboratively solving the task or does not trust neither herself nor the robot. Figure 3 shows the total percentages of choices made by the participants for the emergency scenario. We can observe 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** (as described in Fig. 1). A big majority of participants did not trust the robot to deal with the emergency when tested with **C2**. When tested with **C3** and **C4**, participants chose with similar majorities to solve the task collaboratively and to not trust the robot. The majority of participants preferred to work in collaboration with the robot when tested with **C5**. Summarising, participants chose not to trust the robot when it made severe errors, while they were more inclined to trust in teamwork when the robot made small errors. Moreover, observing the conditions **C3** and **C4** we notice that while the majority of participants chose either to solve the task collaboratively or to not trust the robot, the number of participants who chose to trust the robot increased in **C4**. Therefore, we are inclined to think that participants did not trust the robot more when the severe errors were made by the robot at the beginning of the interaction. 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 ($\phi_c = 0.26, p = 0.001$). We used the adjusted standardised residuals (called Pearson residuals in Agresti [2]) to further analyse the differences between the results obtained. Table 1 shows there is a correlation between the condition **C2** and the choice of the participants to not trust the robot (adjusted value >1.96). We can observe 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

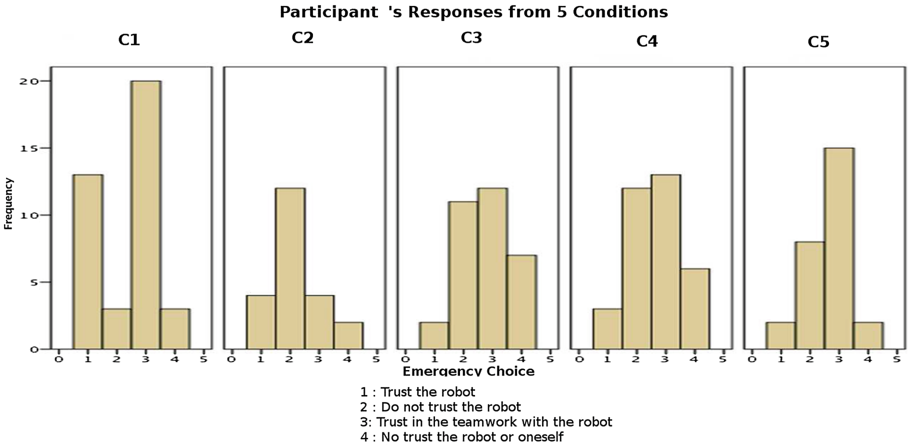


Fig. 3. Responses of participants from different conditions to the Emergency Scenario.

choice 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 were only with one individual, we applied the test only to India and USA. We observed that the association is not statistically significant ($\chi^2(3) = 4.138, p > 0.24$).

Table 1. The adjusted standardised residuals of the Crosstabulation between the choices made by the participants in the emergency scenario and the different conditions presented to the participants.

Condition	Emergency choice			
	Do not trust the robot	Trust the robot	Teamwork with the robot	No trust the robot or oneself
Flawless tasks	−3.5*	3.5*	1.4	−1.1
Big-Big errors	2.7*	0.4	−2.4*	−0.6
Big-Small errors	0.6	−1.6	−0.5	1.7
Small-Big errors	0.8	−1.2	−0.4	0.9
Small-Small errors	0.0	−1.3	1.6	−0.9

5 Discussion and Conclusion

In this study we investigated how human trust depends on the severity and order of presentation of the errors made by a robot. We suggested that there is a correlation between the severity of the error performed by the robot and humans not trusting the robot. We observed the responses of participants of different ages, genders and countries of residence, after interacting with a robot through a storyboard in which their companion robot had erroneous or flawless behaviours. We know that there exist limitations in such online studies. For example the embodiment of a robot plays an important role, but a large percentage of participants (77%) seemed to be truly engaged with the scenarios. We did not find any significant differences in trust tendencies for different ages and different genders of the participants. 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. Our results suggest also that there is a higher tendency to not trust the robot when severe errors happen at the beginning of an interaction. Our findings are also corroborated by Yu et al. [31]'s study. They investigated the correlation between a user's reliance on a system and their trust level. They showed that participants formed their judgements at the beginning of interaction and eventually adjusted it later on, depending on the systems performance. Further investigations will address other open questions. For example, investigating the effect of mechanisms to regain a loss of human trust when the robot has made severe errors.

Acknowledgments. 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. Amazon mechanical turk https://www.mturk.com
2. Agresti, A.: Categorical Data Analysis, 2nd edn. Wiley-Interscience, Chichester, New York (2002)
3. Bainbridge, W.A., Hart, J.W., Kim, E.S., Scassellati, B.: The benefits of interactions with physically present robots over video-displayed agents. Int. J. Social Robot. **3**(1), 41–52 (2011)
4. Billings, D.: Computer poker. University of Alberta M.Sc. thesis (1995)
5. Booth, S., Tompkin, J., Pfister, H., Waldo, J., Gajos, K., Nagpal, R.: Piggybacking robots: human-robot overtrust in university dormitory security, pp. 426–434. ACM (2017)
6. Cameron, D., Aitken, J.M., Collins, E.C., Boorman, L., Chua, A., Fernando, S., McAree, O., Martinez-Hernandez, U., Law, J.: 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)

7. Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., Yanco, H.: Impact of robot failures and feedback on real-time trust. In: ACM/IEEE International Conference on Human-Robot Interaction, pp. 251–258 (2013)
8. Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., Yanco, H.: Effects of changing reliability on trust of robot systems. In: Proceedings of the Seventh Annual ACM IEEE International Conference on Human Robot Interaction, HRI 2012, pp. 73–80 (2012)
9. Deutsch, M.: Trust and suspicion. J. Confl. Resolut. **2**, 265–279 (1958)
10. Golder, S., Donath, J.: Hiding and revealing in online poker games, pp. 370–373 (2004)
11. Gosling, S.D., Rentfrow, P.J., Swann Jr., W.B.: A very brief measure of the big five personality domains. J. Res. Pers. **37**, 504–528 (2003)
12. Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y.C., de Visser, E.J., Parasuraman, R.: A meta-analysis of factors affecting trust in human-robot interaction. Hum. Factors J. Hum. Factors Ergon. Soc. **53**(5), 517–527 (2011)
13. Haselhuhn, M.P., Schweitzer, M.E., Wood, A.M.: How implicit beliefs influence trust recovery. Psychol. Sci. **5**, 645–648 (2010)
14. Koay, K.L., Syrdal, D.S., Walters, M.L., Dautenhahn, K.: 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, pp. 564–569 (2007)
15. Kramer, R.M., Carnevale, P.J.: Trust and intergroup negotiation. In: Brown, R., Gaertner, S.L. (eds.) Handbook of Social Psychology: Intergroup Processes. Blackwell, Boston (2003)
16. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors J. Hum. Factors Ergon. Soc. **46**(1), 50–80 (2004)
17. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manag. Rev. **20**, 709–734 (1995)
18. McKnight, D.H., Choudhury, V., Kacmar, C.: Propensity to trust scale **13**, 339–359 (2001). http://highered.mheducation.com/sites/0073381225/student/view0/chapter7/self-assessment/74.html
19. Muir, B.M., Moray, N.: Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics **39**, 429–460 (1996)
20. Robinette, P., Howard, A.M., Wagner, A.R.: Timing is key for robot trust repair. Social Robotics. LNCS, vol. 9388, pp. 574–583. Springer, Cham (2015). doi:10.1007/978-3-319-25554-5_57
21. Robinette, P., Li, W., Allen, R., Howard, A.M., Wagner, A.R.: Overtrust of robots in emergency evacuation scenarios. In: Proceeding of the Eleventh ACM/IEEE International Conference on Human Robot Interation, HRI 2016, pp. 101–108. IEEE Press, Piscataway (2016)
22. Ross, J.M.: Moderators of trust and reliance across multiple decision aids (Doctoral dissertation), University of Central Florida, Orlando (2008)
23. Rossi, A., Dautenhahn, K., Koay, K.L., Walters, M.L.: Human perceptions of the severity of domestic robot errors. In: Accepted for the Ninth International Conference on Social Robotics, ICSR 2017, 22–24th November 2017, Tsukuba, Japan (2017)
24. Salem, M., Dautenhahn, K.: Evaluating trust and safety in HRI: practical issues and ethical challenges. In: Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2015): Workshop on the Emerging Policy and Ethics of Human-Robot Interaction (2015)

25. Salem, M., Lakatos, G., Amirabdollahian, F., Dautenhahn, K.: Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 141–148 (2015)
26. Schilke, O., Reimann, M., Cook, K.S.: Effect of relationship experience on trust recovery following a breach. Proc. Natl. Acad. Sci. **110**(38), 15236–15241 (2013)
27. Simpson, J.A.: Foundations of interpersonal trust. In: Kruglanski, A.W., Higgins, E.T. (eds.) Social Psychology: Handbook of Basic Principles, pp. 587–607. Guilford, New York (2007)
28. Simpson, J.A.: Psychological foundations of trust. Curr. Dir. Psychol. Sci. **16**(5), 264–268 (2007)
29. Slovic, P.: Perceived risk, trust, and democracy. Risk Anal. **13**, 675–682 (2000)
30. Wang, N., Pynadath, D.V., Unnikrishnan, K.V., Shankar, S., Merchant, C.: Intelligent agents for virtual simulation of human-robot interaction. In: Shumaker, R., Lackey, S. (eds.) VAMR 2015. LNCS, vol. 9179, pp. 228–239. Springer, Cham (2015). doi:10.1007/978-3-319-21067-4_24
31. Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., Chen, F.: User trust dynamics: an investigation driven by differences in system performance, vol. 126745, pp. 307–317. ACM (2017)