

Modeling Trust in Human-Robot Interaction: A Survey

Zahra Rezaei Khavas¹[0000-0002-5268-0197], S. Reza Ahmadzadeh²[0000-0002-6651-8684], and Paul Robinette¹[0000-0001-8066-156X]

¹ Electrical and Computer Engineering, University of Massachusetts Lowell, Lowell, MA 01854, USA Zahra_RezaeiKhavas@student.uml.edu, paul_robinette@uml.edu

² Computer Science, University of Massachusetts Lowell, Lowell, MA 01854, USA
reza_ahmadzadeh@uml.edu

Abstract. As the autonomy and capabilities of robotic systems increase, they are expected to play the role of teammates rather than tools and interact with human collaborators in a more realistic manner, creating a more human-like relationship. Given the impact of trust observed in human-robot interaction (HRI), appropriate trust in robotic collaborators is one of the leading factors influencing the performance of human-robot interaction. Team performance can be diminished if people do not trust robots appropriately by disusing or misusing them based on limited experience. Therefore, trust in HRI needs to be calibrated properly, rather than maximized, to let the formation of an appropriate level of trust in human collaborators. For trust calibration in HRI, trust needs to be modeled first. There are many reviews on factors affecting trust in HRI [22], however, as there are no reviews concentrated on different trust models, in this paper, we review different techniques and methods for trust modeling in HRI. We also present a list of potential directions for further research and some challenges that need to be addressed in future work on human-robot trust modeling.

Keywords: human-robot interaction · human-robot trust · modeling trust in HRI · trust calibration in HRI · trust measurement.

1 Introduction

Trust is one of the essential factors in the development of constructive relationships, including relationships among human individuals and automation. We can mention trust as an overarching concern that affects the effectiveness of a system, especially in terms of safety, performance, and use rate [31]. Having this concern in mind, trust has become a critical element in the design and development of automated systems [35]. Autonomous systems are being designed and developed with increased levels of independence and decision-making capabilities, and these capacities will be efficient in uncertain situations [55].

Human-robot trust is an important branch of Human-Robot Interaction (HRI), which has recently gained increasing attention among scholars in many

disciplines, such as Computer Engineering [49], Psychology [22], Computer Science [10] and Mechanical Engineering [52]. Trust is a significant factor that needs to be taken into consideration when robots are going to work as teammates in human-robot teams [19], used as autonomous agents [57], or where robots are going to be used in a complex and dangerous situations [42,49]. In many cases, trust is the main factor determining how much a robotic agent would be accepted and used by the human [44].

Weak partnerships resulting from inappropriate or non-calibrated trust between humans and robots might cause misuse or disuse of a robotic agent. Misuse refers to the failures that occur due to a user's over-trust of the robotic agent (e.g., accepting all the solutions and results presented by the robot without questioning). In contrast, disuse refers to the failures that occur due to human under-trust to the robotic agent (e.g., rejecting the capabilities of a robotic agent) [31]. To prevent human operators' misuse and disuse of robots, trust needs to be calibrated. We need to model trust to generate a measure of it and to enable proper trust calibration [43].

There are many reviews available in HRI which are mostly concentrated on factors affecting trust [22,55], but none of them is concentrated on modeling trust and different trust models in HRI. This document's purpose is to review different studies concentrated on trust modeling in HRI. This review starts by defining trust, followed by summarizing factors affecting trust, reviewing studies focused on modeling trust, and a summary of input and output elements to the trust models. We conclude with a discussion of the shortcomings and challenges of trust modeling and future avenues of research.

2 Definition of Trust in HRI

According to psychologists, trust is a mental state of a human [2]. Numerous researchers have extensively explored the notion of trust for decades. Trust is not limited to just interpersonal interactions. It underlies different forms of interactions, such as banks' interaction with customers, governments with citizens, employers with employees, etc., [61]. Therefore, we can say trust can affect human-robot interaction, as it can affect a human user or collaborator's willingness to assign tasks, share information, cooperate, provide support, accept results, and interact with a robot [16].

As trust is one of the necessities for building a successful human-robot interaction, we need to create methodologies to model, measure, and calibrate trust. The first step for modeling trust is a clear definition of trust. However, despite the broad efforts and number of studies concentrating on trust, there is little consensus on a single definition, as the definition of trust is heavily dependent on the context in which trust is being discussed [6]. This is highlighted, for instance, in [41], where trust is defined as "*the reliance by one agent that actions prejudicial to the well-being of that agent will not be undertaken by influential others*". Around each application and domain that robots are used, trust needs to be defined, measured and explored explicitly. For example, in high-risk

robotic applications such as emergency evacuation robots [49], the definition of trust might differ substantially from low-risk robotic applications such as border tracking robots [65]. One of the most thorough definitions of trust, which is deployed by many other studies concentrated on human-robot trust, is by Lee and See [31]. They define trust from the perspective of automation. This definition was generated by reviewing many other studies concentrated on defining trust and was complementary for many other works. They define trust as: “*the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability*”. This definition of trust is accepted and used by many studies on trust in HRI. Wagner et al. [61] also provided a comprehensive definition for trust: “*a belief, held by the trustor, that the trustee will act in a manner that mitigates the trustor’s risk in a situation in which the trustor has put its outcomes at risk*”. They also provided a model for determining if an interaction demands trust or not. All these definitions have one thing in common, that is: “whether robot’s actions and behaviors correspond to human’s interest or not?” To address this concern in each robotic application trust needs to be modeled based on human interests in that domain.

3 Factors Affecting Trust

Studies on factors affecting trust in HRI can be considered an extension to the studies on factors affecting human-automation trust. In [31], the authors review the factors affecting trust in humans and generalize these factors to factors affecting trust in automation. Many other studies also review and analyze factors affecting trust in human-automation Interaction (HAI) [23,55]. However, robots differ from other forms of automation in many cases, such as mobility, embodiment, and unfamiliarity to the general public. Therefore, factors affecting trust in HRI need to be investigated separately.

Hancock et al. [22] provides a meta-analysis of factors affecting trust in HRI and classifies these factors in three categories each consist of two subcategories: 1. Human-related factors (i.e., including ability-based factors, characteristics); 2. Robot-related factors (i.e., including performance-based factors, attribute-based factors); and 3. Environmental factors (i.e., including team collaboration, tasking). In this study, we review and classify recent studies on factors affecting trust with a similar classification basis as in [22] with some updates in categories and subcategories. We classify factors affecting trust in HRI into three categories: 1. Robot-related factors (including robot-performance, robot-appearance, and robot-behaviors), 2. Human-related factors and 3. Task and Environment-related factors. Table 1 shows our classification of factors affecting trust. Most of these factors are similar to the ones that are mentioned in Hancock et al. [22]. However, some factors are excluded as recent studies paid less attentions to them and some new factors that are stressed more in recent studies are added to this table. Some factors are also classified under different categories here in this work compared to [22].

1. **Robot-Related Factors:** Robot-related factors have the greatest effect on the trust in HRI [22]. This justifies the great number of studies concentrated on robot-related factors. We classify robot-related factors under three sub-categories:
 - (a) Performance-related factors: These factors determine the quality of an operation performed by the robot from the human operator's point of view. Of these factors we can mention *reliability*, *faulty behavior*, *frequency of fault occurrence* [12], *timing of error* [33], *transparency*, *feedback* [53,39], *level of situation awareness* [5], *false alarms* [69], and *level of autonomy* [28].
 - (b) Behavior-related factors: Advancements in robotic systems in recent years have caused people to consider them more like teammates than tools. Increased autonomy of robots caused an increase in the perceived intelligence of robots by humans. These advancements altered the form of HRI to a more naturalistic interaction [43]. Approaching a more human-like interaction with robots cause people to consider the intention of a robot's behavior. Some general behaviors of the robot such as *likeability* (e.g., gaze behaviors and greeting) [38], *proximity* (e.g., physical and physiological proximity) [38,40,62], *engagement* [50], *confess to the reliability* [63,39] and *harmony of robot personality with the task* (e.g., introverted robotic security guard and an extroverted robotic nurse) [59] can affect formation and maintenance of trust. There are also some specific behaviors of robots that can affect trust repair after failure or wrong and misleading action such as *apology*, *making excuses* or *explanations* and *dialogues* [48,33,58,56,39].
 - (c) Appearance-related factors: People consider the personality of a robot based on the robot's appearance and behavior during interactions. Some features in robot appearance such as *anthropomorphism* [39], *gender* [59], *harmony of task with robot's appearance* [50], and *similarity with human collaborator* [70] affect trust in HRI.
2. **Human-Related Factors:** Although, according to [22], human-related factors have the least effect on trust, they are important in the formation and fluctuation of trust during human-robot interaction. Many studies investigated the effect of these factors on trust. Of these factors, individual's *gender*, *subjective feeling* toward robots [40], *initial expectations* of people toward automation [69], *previous experience* of the individuals with robots [62], *culture* [32] and also the human's *understanding of the system* [43] are some of the human-related factors that are addressed in different studies.
3. **Task and Environment-Related Factors:** Based on Hancock et al. [22], task and environment-related factors have the second greatest effect on human-robot trust. Many factors related to the the task type and location such as *rationality* and *revocability* of the tasks [53], *risk* and *human-safety* [49], *workload* [12], *nature of the task* (e.g., nurse or security robot) [59], *task duration* [62,69], *in-group membership* [19,46], *physical presence of robot* [3,39] and *task site* [28] are thoroughly investigated by researchers in HRI.

Table 1. Factors affecting development of trust in human-robot interaction

1. Robot-Related		3. Task & Environment-Related
(a) Performance-Related	(c) Appearance-Related	Nature of task [59,28]
Dependability, reliability and error [12,11,53,49,39,33]	Similarity with operator [70]	Physical presence of robot in task site [3,39]
Level of [28,18]	Gender [70,59]	In-group membership [46,19]
Situation awareness, feedback and Transparency [11,39,5,63]	Harmony of appearance with task [50]	Duration of interaction [69,62]
(b) Behavior-Related	Anthropomorphism [39]	Task site [28]
Dialogues [33]	2. Human-Related	Revocability [53]
Proximity [38,40,62]	Personality [53]	Rationality [53,49,3]
Likeability and friendliness [38]	Culture [32]	Risk [49,28,70]
Personality (harmony with task) [59]	Understanding of the system [43]	Workload, complexity and required level of multi tasking [12]
Confess to reliability [63,39]	Demographics [40,34]	
Apology for failure [58,56,39]	Subjective feeling [40,69]	
Engagement [50]	Experience with robots [34]	

4 Current Research on Trust Modeling

Trust in HRI has a lot in common with trust in HAI, which has been studied at length. Muir et al. [37] found the available definitions for trust between humans inconsistent with the nature of HAI based on the multidimensional construct of trust. She defined a trust models for HAI which was based on model of human expectation of automation proposed by Barber et al. [4]. Lee and Moray [29] built upon Muir’s strategy for modeling trust, which was identifying independent variables that influence trust, and introduced another model for trust. Later, other researchers have modeled the operator’s trust in automation, considering more factors affecting trust [24,14,31]. These models were finally classified into five groups [36]: regression-based models [37,29], time series models [30], qualitative models [47], argument based probabilistic models [9], and neural net models [14].

There are many similarities between trust in HRI and trust in HAI. However, most of the models generated for modeling trust in HAI are inconsistent with the needs of HRI. According to Desai [13] “*these models do not consider some factors that appear while working with robots such as situational awareness, usability of the interface, physical presence of robots (co-located with human or remote-located), limitations and complexities of the operating environment, workload, task difficulty, etc. which influence HRI considerably*”. Desai et al. [13] introduced a schematic of a model considering some factors affecting human-robot trust in conjugate with factors affecting human-automation trust. Yagoda et al. [68] introduced one of the very first models for human-robot trust based on the different dimensions of a human-robot interaction task and validity assessment of each of these directions by subject matter experts (SMEs) in the HRI. Desai et al. [10] also was one of the pioneers in modeling trust in HRI. They generated a more detailed model for trust in human and autonomous robot tele-operation. This model used the Area Under Trust Curve (AUTC) measure to account for an individual’s entire interactive experience with the robot.

According to several studies, there is a strong correlation between the level of trust in human-robot teammates with the performance of the robotic agent’s work, and it also impacts their interaction quality [31,10]. According to Xu [66], high levels of trust among human-robot teammates often demonstrate great synergy, in which matched decision-making capabilities of the human member in the team complements the exhaustive controlling and executing capabilities of the robotic agent. In contrast, a low level of trust among human-robot teammates might cause a human to refuse to delegate tasks to the robotic agent or sometimes decide to disable the robotic agent [66]. Since there is often a high correlation among trust and performance of the work in human-robot collaboration, trust has been modeled based on the performance in many studies [64,52]. Most of these performance-based trust models are used as a feedback loop to adjust actual performance of the robot to the human’s expectation of its performance to convince a human to act trusting toward robot. There are also some trust models based on the performance of robot operations, which are not aimed to modify the performance of collaboration. For example, in [45] a performance-based trust model for multi-robot tasks is designed to detect robotic agents that are not reliable. Then, less reliable agents are assigned to do less critical tasks or sometimes disregarded while assigning tasks. On the other hand, the human-robot collaboration’s performance can also be modeled based on trust [16,17]. These models are used to modify the robot’s trust-related behavior to manage the overall performance of the collaboration.

A prevalent class of human-robot collaborations is supervisory collaboration in which the human plays the role of supervisor and the robot plays the role of worker. The supervisor delegates tasks to the worker and oversees the performance of the operation. The supervisor also has the authority to take control of the robot when the robot is doing something wrong. The model of trust for supervisory collaboration presented in [64] is based on trust in between-human collaboration. It generates a quantity showing the compatibility of robot performance with the human expectation, enabling the robot to modify its performance to fulfill human expectations and improve trust. Later this trust model was improved [67], and more factors affecting trust in supervisory collaboration such as failure rate in the autonomous agent and the rate of supervisor intervention were involved in designing the trust model. The Online Probabilistic Trust Inference Model (OPTIMo) [65] is another model of trust in the supervisory collaboration introduced by the same research group. This model formulates Bayesian beliefs over the human’s trust status based on the robot’s performance on the task over time to generate a real-time estimate of the human’s trust.

When trust can be modeled and measured in a real-time manner in human-robot collaboration, it can help the robot repair trust whenever the human starts under-trusting the robot [65]. A real-time model of trust (trust-POMDP) for human-robot peer-to-peer collaboration is introduced in [8], which integrates measured trust in the robot’s decision-making. The trust-POMDP model closes the loop between measured trust by the real-time trust model and robot decision-making process to maximize collaboration performance. This model grants a

robot the ability to influence human trust systematically to reduce and increase trust in over-reliance and under-reliance situations, respectively.

Hancock et al. [22] provided a meta-analysis of a great number of factors affecting trust in HRI, and quantitative measurement for the effect of human-based, robot-based, and environment-based factors on trust. Later they developed a model of human-robot team trust based on their findings in the meta-analysis [22]. In addition to the effect of three classes of factors affecting trust introduced in the prior work, they considered the effects of training and design implications on the final model of trust [54].

Some recent studies in trust modeling use subjective trust measurement techniques in companions with objective trust measurement techniques. These techniques are deployed in both HAI and HRI for increasing the accuracy and robustness of trust measurements. There are some studies in human-automation trust, human-computer interaction, and human trust in artificial intelligence that use psycho-physiological measurements for trust modeling [1,20,21]. Khalid et al. [27] also introduce a trust model in HRI, which uses facial expressions, voice features, and extracted heart rate features in combination with the self-reported trust of humans to model trust. This trust model classifies the trust level into one of the low, natural, and high trust levels using a Neuro-fuzzy trust classifier.

5 Trust Model: Inputs and Outputs

Trust models formulate the effect of factors on the formation and variation of trust in robots. In fact, trust models use factors affecting trust to estimate trust. Since these factors vary in different domains and environments, input factors to the trust models vary based on the application domain. For example, Robinette et al. [51] models trust in emergency evacuation based on the situational risk (e.g., amount of danger perceived by the human in the environment around him) and agent risk (e.g., agent's behavior and appearance) to model perceived trust by the human and the human's decision to trust the robot's guidance or not. In contrast, [64] proposes a trust model for a supervisory collaboration and formulates trust as a function of the robot's success and failure in performing the task. The output of this trust model is closing the loop between human trust and robot function by adjusting the robot's action to improve the collaboration efficiency. Finally, [68] proposes a more general trust model based on team configuration, task, system, context, and team process to scale trust for trust measurement.

Many of the studies on modeling trust in HRI consider the performance of collaboration as one of the main input elements for their model [65,17,52]. Most of these models consider the effect of performance in conjugate with some other factors. For instance, the OPTIMO probabilistic trust model [65] uses rates of robot's failures, and human interventions in conjugate with task performance as inputs to the model to estimate the human's degree of trust in a robotic teammate. Meanwhile, [17] uses the operator's perception of system capabilities, past experience, and training to assess initial trust. Trust gets updated in a loop based on system performance, cognitive workload, and frequency of

changes from tele-operation to autonomous operation. This trust model's output is a measure of gain and loss of trust and the impact of these changes of trust on collaboration performance. Sadrifaridpour et al. [52] models trust based on human performance (i.e., muscles fatigue and dynamics of recovery), robot performance (i.e., speed of robot doing the specific task), workload and human expectation of task performance. The output of this model is feedback to the robot to adjust its performance according to operator desires.

6 Conclusion and Future Work

Most of the existing trust models in HRI are developed for a specific form of human-robot interaction or a specific type of robotic agent. For example, in [51], a model of trust is specified for evacuation robots; in [12] the trust model is specified on robots with shared control, and in [65] the generated model is usable for supervisory collaboration. As each of these trust models belongs to a specific domain, they can not be compared with each other, and there is no scale to evaluate their accuracy. A general model of trust is needed in HRI to evaluate the accuracy of other trust models. Such a model can be applied to any robotic domain and even can be deployed by newly emerged robotic areas and eliminate the need for new trust models for these areas.

Trust is a subject of interest for research in many other fields such as psychology, sociology, and even physiology. In these fields, other indicators of trust are used for trust measurement. For example, some studies use physiological indicators, such as oxytocin-related measures [60,15,25], and objective measures, such as trust games that assess actual investment behavior [7,26]. These trust assessment methods can be used in HRI to develop a trust model that is independent of the countless parameters that affect trust.

Many studies examined the effect of robot failure on trust [49,12,53,39]. However, there are limited studies focused on modeling the effect of robot failure on trust. While Desai et al. [10] investigated the effect of robot's reduced performance and timing of performance reduction on the operator's trust in a tele-operation human-robot interaction, the effect of trust violation needs to be investigated more deeply. Different forms of trust violation in different robotic domains need to be explored. A variety of factors affecting trust-loss and trust-repair after trust violation, such as task type, risk, robot type, robot behavior, etc. need to be taken into account in modeling the effect of robot failure on trust. Modeling fluctuations of trust in a human after a robot failure would help estimate the time required for human trust to tend to a steady-state and formulate timing of trust repair after trust violation.

Both modeling trust and modeling the effects of failure on trust need to be explored in a more general context. Since there are many different factors affecting trust, it will be challenging to develop a conclusive trust model that incorporates all of these factors. Therefore, future research into trust modeling in HRI needs to be more focused on developing general models of trust, which are based on measures other than the countless factors affecting trust. Such models

would not be affected by the emergence of new factors affecting trust in existing or new robotic domains and eliminate the need to develop new trust models for new domains of the ever-evolving world of robotics.

References

1. Ajenaghughrure, I.B., Sousa, S.C., Kosunen, I.J., Lamas, D.: Predictive model to assess user trust: a psycho-physiological approach. In: Proceedings of the 10th Indian Conf on Human-Computer Interaction. pp. 1–10 (2019)
2. Atoyan, H., Duquet, J.R., Robert, J.M.: Trust in new decision aid systems. In: Proceedings of the 18th Conf on l’Interaction Homme-Machine. pp. 115–122 (2006)
3. Bainbridge, W.A., Hart, J.W., Kim, E.S., Scassellati, B.: The benefits of interactions with physically present robots over video-displayed agents. *Intl Journal of Social Robotics* **3**(1), 41–52 (2011)
4. Barber, B.: The logic and limits of trust (1983)
5. Boyce, M.W., Chen, J.Y., Selkowitz, A.R., Lakhmani, S.G.: Effects of agent transparency on operator trust. In: Proceedings of the Tenth Annual ACM/IEEE Intl Conf on HRI Extended Abstracts. pp. 179–180 (2015)
6. Cameron, D., Aitken, J., Collins, E., 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 (2015)
7. Chang, L.J., Doll, B.B., van’t Wout, M., Frank, M.J., Sanfey, A.G.: Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive psy* **61**(2), 87–105 (2010)
8. Chen, M., Nikolaidis, S., Soh, H., Hsu, D., Srinivasa, S.: Planning with trust for human-robot collaboration. In: Proceedings of the 2018 ACM/IEEE Intl Conf on HRI. pp. 307–315 (2018)
9. Cohen, M.S., Parasuraman, R., Freeman, J.T.: Trust in decision aids: A model and its training implications. In: in Proc. Command and Control Research and Technology Symp. Citeseer (1998)
10. Desai, M.: Modeling trust to improve human-robot interaction. Ph.D. thesis, University of Massachusetts Lowell (2012)
11. Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., Yanco, H.: Impact of robot failures and feedback on real-time trust. In: 2013 8th ACM/IEEE Intl Conf on HRI). pp. 251–258. IEEE (2013)
12. 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: 2012 7th ACM/IEEE Intl Conf on HRI. pp. 73–80. IEEE (2012)
13. Desai, M., Stubbs, K., Steinfeld, A., Yanco, H.: Creating trustworthy robots: Lessons and inspirations from automated systems (2009)
14. Farrell, S., Lewandowsky, S.: A connectionist model of complacency and adaptive recovery under automation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **26**(2), 395 (2000)
15. Ferreira, F., Lopes da Costa, R., Pereira, L., Jerónimo, C., Dias, Á.: The relationship between chemical of happiness, chemical of stress, leadership, motivation and organizational trust: A case study on brazilian workers. The relationship between chemical of happiness, chemical of stress, leadership, motivation and organizational trust: a case study on Brazilian workers (2), 89–100 (2018)
16. Freedy, A., DeVisser, E., Weltman, G., Coeyman, N.: Measurement of trust in human-robot collaboration. In: 2007 Intl Symposium on Collaborative Technologies and Systems. pp. 106–114. IEEE (2007)

17. Gao, F., Clare, A., Macbeth, J., Cummings, M.: Modeling the impact of operator trust on performance in multiple robot control. In: 2013 AAAI Spring Symposium Series (2013)
18. Goodrich, M.A., Olsen, D.R., Crandall, J.W., Palmer, T.J.: Experiments in adjustable autonomy. In: Proceedings of IJCAI Workshop on autonomy, delegation and control: interacting with intelligent agents. pp. 1624–1629. Seattle, WA (2001)
19. Groom, V., Nass, C.: Can robots be teammates?: Benchmarks in human-robot teams. *Interaction Studies* **8**(3), 483–500 (2007)
20. Gulati, S., Sousa, S., Lamas, D.: Modelling trust: An empirical assessment. In: IFIP Conf on Human-Computer Interaction. pp. 40–61. Springer (2017)
21. Gulati, S., Sousa, S., Lamas, D.: Design, development and evaluation of a human-computer trust scale. *Behaviour & Info Tech* **38**(10), 1004–1015 (2019)
22. Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y., De Visser, E.J., Parasuraman, R.: A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* **53**(5), 517–527 (2011)
23. Hoff, K.A., Bashir, M.: Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* **57**(3), 407–434 (2015)
24. Itoh, M., Tanaka, K.: Mathematical modeling of trust in automation: Trust, distrust, and mistrust. In: Proceedings of the human factors and ergonomics society annual meeting. vol. 44, pp. 9–12. SAGE Publications Sage LA, CA (2000)
25. Johnson, N.D., Mislin, A.A.: Trust games: A meta-analysis. *Journal of Economic Psychology* **32**(5), 865–889 (2011)
26. Keri, S., Kiss, I., Kelemen, O.: Sharing secrets: oxytocin and trust in schizophrenia. *Social neuroscience* **4**(4), 287–293 (2009)
27. Khalid, H.M., Shiung, L.W., Nooralishahi, P., Rasool, Z., Helander, M.G., Kiong, L.C., Ai-vyrn, C.: Exploring psycho-physiological correlates to trust: Implications for human-robot-human interaction. In: Proceedings of the human factors and ergonomics society annual meeting. vol. 60, pp. 697–701. SAGE Publications Sage CA: Los Angeles, CA (2016)
28. Lazanyi, K., Maraczi, G.: Dispositional trust—do we trust autonomous cars? In: 2017 IEEE 15th Intl Symposium on Intelligent Systems and Informatics (SISY). pp. 000135–000140. IEEE (2017)
29. Lee, J., Moray, N.: Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* **35**(10), 1243–1270 (1992)
30. Lee, J.D., Moray, N.: Trust, self-confidence, and operators’ adaptation to automation. *Intl journal of human-computer studies* **40**(1), 153–184 (1994)
31. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. *Human factors* **46**(1), 50–80 (2004)
32. Li, D., Rau, P.P., Li, Y.: A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics* **2**(2), 175–186 (2010)
33. Lucas, G.M., Boberg, J., Traum, D., Artstein, R., Gratch, J., Gainer, A., Johnson, E., Leuski, A., Nakano, M.: Getting to know each other: The role of social dialogue in recovery from errors in social robots. In: Proceedings of the 2018 acm/ieee Intl Conf on HRI. pp. 344–351 (2018)
34. M Tsui, K., Desai, M., A Yanco, H., Cramer, H., Kemper, N.: Measuring attitudes towards telepresence robots. *Intl journal of intelligent control and systems* **16** (2011)
35. Martelaro, N., Nneji, V.C., Ju, W., Hinds, P.: Tell me more designing hri to encourage more trust, disclosure, and companionship. In: 2016 11th ACM/IEEE Intl Conf on HRI. pp. 181–188. IEEE (2016)

36. Moray, N., Inagaki, T.: Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control* **21**(4-5), 203–211 (1999)
37. Muir, B.M.: Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* **37**(11), 1905–1922 (1994)
38. Mumm, J., Mutlu, B.: Human-robot proxemics: physical and psychological distancing in human-robot interaction. In: *Proceedings of the 6th Intl Conf on HRI*. pp. 331–338 (2011)
39. Natarajan, M., Gombolay, M.: Effects of anthropomorphism and accountability on trust in human robot interaction. In: *Proceedings of the 2020 ACM/IEEE Intl Conf on HRI*. pp. 33–42 (2020)
40. Obaid, M., Sandoval, E.B., Zlotowski, J., Moltchanova, E., Basedow, C.A., Bartneck, C.: Stop! that is close enough. how body postures influence human-robot proximity. In: *2016 25th IEEE Intl Symposium on Robot and Human Interactive Communication (RO-MAN)*. pp. 354–361. IEEE (2016)
41. Oleson, K.E., Billings, D.R., Kocsis, V., Chen, J.Y., Hancock, P.A.: Antecedents of trust in human-robot collaborations. In: *2011 IEEE Intl Multi-Disciplinary Conf on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. pp. 175–178. IEEE (2011)
42. Ososky, S., Sanders, T., Jentsch, F., Hancock, P., Chen, J.Y.: Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In: *Unmanned Systems Technology XVI*. vol. 9084, p. 90840E. Intl Society for Optics and Photonics (2014)
43. Ososky, S., Schuster, D., Phillips, E., Jentsch, F.G.: Building appropriate trust in human-robot teams. In: *2013 AAAI Spring Symposium Series* (2013)
44. Parasuraman, R., Riley, V.: Humans and automation: Use, misuse, disuse, abuse. *Human factors* **39**(2), 230–253 (1997)
45. Pippin, C., Christensen, H.: Trust modeling in multi-robot patrolling. In: *2014 IEEE Intl Conf on Robotics and Automation (ICRA)*. pp. 59–66. IEEE (2014)
46. Rau, P.P., Li, Y., Li, D.: Effects of communication style and culture on ability to accept recommendations from robots. *Computers in Human Behavior* **25**(2), 587–595 (2009)
47. Riley, V.: Operator reliance on automation: Theory and data. *Automation and human performance: Theory and applications* pp. 19–35 (1996)
48. Robinette, P., Howard, A.M., Wagner, A.R.: Timing is key for robot trust repair. In: *Intl Conf on social robotics*. pp. 574–583. Springer (2015)
49. Robinette, P., Li, W., Allen, R., Howard, A.M., Wagner, A.R.: Overtrust of robots in emergency evacuation scenarios. In: *2016 11th ACM/IEEE Intl Conf on HRI*. pp. 101–108. IEEE (2016)
50. Robinette, P., Wagner, A.R., Howard, A.M.: Building and maintaining trust between humans and guidance robots in an emergency. In: *2013 AAAI Spring Symposium Series* (2013)
51. Robinette, P., Wagner, A.R., Howard, A.M.: Modeling human-robot trust in emergencies. In: *2014 AAAI Spring Symposium Series* (2014)
52. Sadrfaridpour, B., Saeidi, H., Burke, J., Madathil, K., Wang, Y.: Modeling and control of trust in human-robot collaborative manufacturing. In: *Robust Intelligence and Trust in Autonomous Systems*, pp. 115–141. Springer (2016)
53. 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: *2015 10th ACM/Intl Conf on HRI*. pp. 1–8. IEEE (2015)

54. Sanders, T., Oleson, K.E., Billings, D.R., Chen, J.Y., Hancock, P.A.: A model of human-robot trust: Theoretical model development. In: *Proceedings of the human factors and ergonomics society annual meeting*. vol. 55, pp. 1432–1436. SAGE Publications Sage CA: Los Angeles, CA (2011)
55. Schaefer, K.E., Chen, J.Y., Szalma, J.L., Hancock, P.A.: A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors* **58**(3), 377–400 (2016)
56. Sebo, S.S., Krishnamurthi, P., Scassellati, B.: “i don’t believe you”: Investigating the effects of robot trust violation and repair. In: *2019 14th ACM/IEEE Intl Conf on HRI*. pp. 57–65. IEEE (2019)
57. Selkowitz, A., Lakhmani, S., Chen, J.Y., Boyce, M.: The effects of agent transparency on human interaction with an autonomous robotic agent. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. vol. 59, pp. 806–810. SAGE Publications Sage CA: Los Angeles, CA (2015)
58. Strohkorb Sebo, S., Traeger, M., Jung, M., Scassellati, B.: The ripple effects of vulnerability: The effects of a robot’s vulnerable behavior on trust in human-robot teams. In: *Proceedings of the 2018 ACM/IEEE Intl Conf on HRI*. pp. 178–186 (2018)
59. Tay, B., Jung, Y., Park, T.: When stereotypes meet robots: the double-edge sword of robot gender and personality in human-robot interaction. *Computers in Human Behavior* **38**, 75–84 (2014)
60. Uvnas-Moberg, K., Petersson, M.: Oxytocin, a mediator of anti-stress, well-being, social interaction, growth and healing. *Z Psychosom Med Psychother* **51**(1), 57–80 (2005)
61. Wagner, A.R., Arkin, R.C.: Recognizing situations that demand trust. In: *2011 RO-MAN*. pp. 7–14. IEEE (2011)
62. Walters, M.L., Oskoei, M.A., Syrdal, D.S., Dautenhahn, K.: A long-term human-robot proxemic study. In: *2011 RO-MAN*. pp. 137–142. IEEE (2011)
63. Wang, N., Pynadath, D.V., Hill, S.G.: Trust calibration within a human-robot team: Comparing automatically generated explanations. In: *2016 11th ACM/IEEE International Conf on HRI*. pp. 109–116. IEEE (2016)
64. Xu, A., Dudek, G.: Trust-driven interactive visual navigation for autonomous robots. In: *2012 IEEE Intl Conf on Robotics and Automation*. pp. 3922–3929. IEEE (2012)
65. Xu, A., Dudek, G.: Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In: *2015 10th ACM/IEEE Intl Conf on HRI*. pp. 221–228. IEEE (2015)
66. Xu, A., Dudek, G.: Maintaining efficient collaboration with trust-seeking robots. In: *2016 IEEE/RSJ Intl Conf on Intelligent Robots and Systems (IROS)*. pp. 3312–3319. IEEE (2016)
67. Xu, A., Dudek, G.: Towards modeling real-time trust in asymmetric human-robot collaborations. In: *Robotics Research*, pp. 113–129. Springer (2016)
68. Yagoda, R.E., Gillan, D.J.: You want me to trust a robot? the development of a human-robot interaction trust scale. *Intl Journal of Social Robotics* **4**(3), 235–248 (2012)
69. Yang, X.J., Unhelkar, V.V., Li, K., Shah, J.A.: Evaluating effects of user experience and system transparency on trust in automation. In: *2017 12th ACM/IEEE Intl Conf on HRI*. pp. 408–416. IEEE (2017)
70. You, S., Robert Jr, L.P.: Human-robot similarity and willingness to work with a robotic co-worker. In: *Proceedings of the 2018 ACM/IEEE Intl Conf on HRI*. pp. 251–260 (2018)