An Exploration of Trust in Human-Robot Interaction: From Measurement to Repair Strategies and Design Principles

Fatima Ayoub[1], Aphra Kerr[2] and Rudi Villing[1]

[1] Maynooth University, Maynooth, Ireland
[2] University College Dublin, Dublin, Ireland
fatima.ayoub.2020@mumail.ie

Abstract. This paper presents an interdisciplinary scoping review of literature on trust and trust repair in human-robot interaction (HRI), focusing on social and service robots. The primary aim is to provide a comprehensive analysis of trust in social robotics and the methodologies for measuring trust within HRI. Distinct from prior reviews, this work delves into both trust models and trust measurement methods highlighting the complex nature of trust. This study assesses trust repair strategies, including promises and explanations, in different contexts. It also investigates scenarios of compromised trust and explores both communicative and proactive repair approaches. Finally, the paper presents nine key design principles such as safety, transparency, apology mechanism etc., derived from the reviewed literature, as guidelines for the development of social and service robots. These principles provide researchers with a path for creating robots that can build and maintain trust, particularly when they are making mistakes and need to correct them.

Keywords: Trust in HRI, Trust models, Trust measurement, Trust repair, Trust principles.

1 Introduction

Trust is a complex and multidimensional concept that can be defined as the willingness to accept vulnerability or risk based on expectations of the intentions or behaviors of another human or non-human entity [1]. Trust is essential for the successful adoption and integration of social and service robots into various domains of society, such as education, healthcare, and entertainment. Social robots are autonomous robots that interact and communicate with humans or other autonomous physical agents by following social behaviors and rules attached to their role [2]. Service robots, as defined by the International Organization for Standardization, are robots that perform useful tasks for humans or equipment excluding industrial automation applications [3]. Trust affects how humans perceive, interact with, and rely on these robots, and influences their acceptance and satisfaction with these technologies. Trust also forms a vital element for the effective functioning of these systems, as it enables coordination, cooperation, and collaboration between humans and robots [4].

In the study of trust within robotics, it is important to differentiate between the user's state of trust and a robot's trustworthiness. Trustworthiness is an inherent characteristic

of the robot, defined by its ability to consistently demonstrate reliability, skillfulness, and the capacity to fulfill user expectations across different situations [5]. For example, a robot's trustworthiness is assessed through its steady performance, interaction precision, and commitment to safety standards [6]. In social contexts, this also includes the robot's display of social signals and conduct that resonate with human norms and ethics. Furthermore, trust and trustworthiness in robots are fundamental elements to their acceptance and integration within society. These qualities influence user perceptions and their readiness to interact with robots. This is especially critical in situations where tasks involve some level of risk or require significant reliance on the robot's behavior [7].

However, a user's state of trust is not a static or fixed attribute, but rather a dynamic and context-dependent process that can change over time and across situations [8]. Therefore, it is important to understand how trust can be built, measured, and repaired in HRI, especially when trust violations occur due to robot errors, failures, or misbehaviors. Trust violations can have negative consequences for the human-robot relationship, such as reduced trust, satisfaction, performance, willingness to interact, and sometimes complete rejection of a technology. The concept of "trust repair" in robotics literature was first discussed in 2015 [9]. Trust repair in social robotics refers to the process of recovering or improving the level of trust that a human has in a social robot, after the trust has been violated [10].

While there is extensive literature on human-to-human trust and trust in HRI (see surveys and reviews [7, 11–15]), this paper presents an interdisciplinary scoping review of the literature on trust in robots (social and service) and trust repair in HRI. Unlike previous reviews, this study reviews the trust models, the methods for measuring trust, and the strategies for repairing trust. Trust repair is an important aspect of human-robot collaboration, as robots are prone to errors and failures in real-world scenarios. Therefore, this paper reviews the effectiveness of different trust repair techniques, such as promises and explanations, and how they vary depending on the context of the interaction. In addition to our main objectives, we also identify the essential principles that are important for building trust. These principles are derived from our review of the literature and can be helpful when designing social robots.

This paper has the following structure: In Section 2, we describe the criteria for choosing the papers and the method we used to conduct the review. In Section 3, we identify the techniques to build and assess trust and also address the topic of trust failure and trust repair. Section 4 presents the design guidelines that one should follow to ensure the trustworthiness of a social robot. We present our conclusions in the final section.

2 Methodology

We conducted a comprehensive interdisciplinary scoping review of published research within the domains of trust focusing on trust in social and service robotics. This established methodology was employed to identify key factors and characteristics within our field of interest. Additionally, the methodology helps to find research gaps and to map the general research patterns in the field [16, 17]. This is especially helpful in a

developing field such as ours, where it is important to build a basic understanding of the topic by bringing together different areas of research.

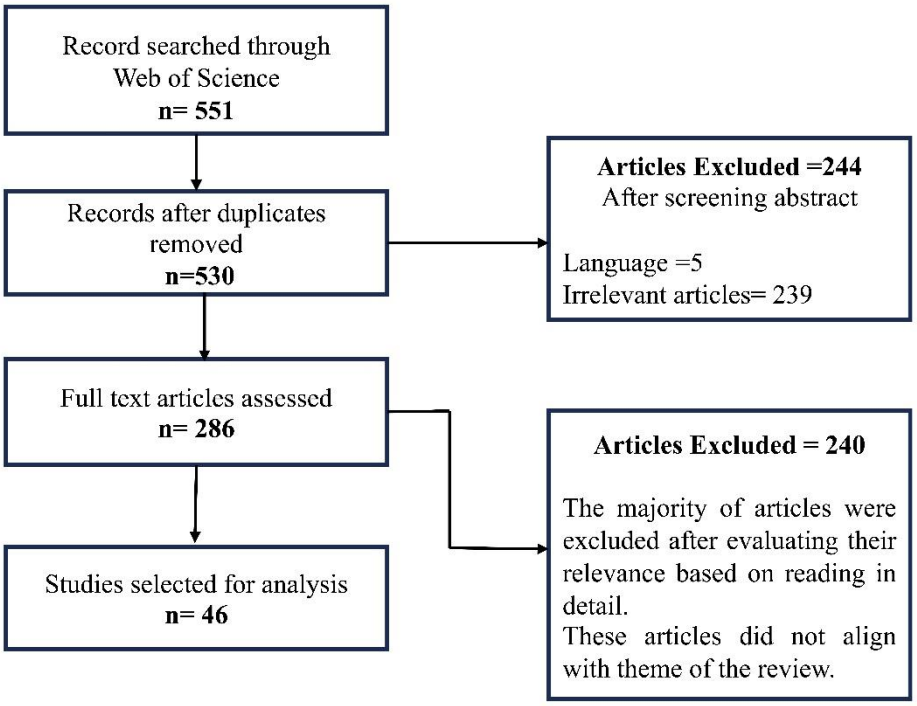

Fig. 1. The process of literature selection for analysis, from the initial identification to the final selection, with all reasons for exclusion.

A comprehensive literature search was conducted on the Web of Science, encompassing conference and peer-reviewed journal articles. The search, carried out in February 2024 for past 5 years and used the search terms “Trust in robotics”, “Trust” AND “Robotics”, “Trust measurement in robotics”, “Trust measurement” AND “Robotics” “Trust repair in robotics”, and “Trust repair” AND “Robotics” to get initial papers. The search revealed 551 articles on robotics and trust repair. From the initial 551 articles on trust in robotics and repair, a review was conducted on 46 pertinent studies from fields including robotics, computer science, psychology, and sociology, all centered on trust in robots from the user’s perspective. The process of exclusion is illustrated in Fig. 1.

Articles were excluded based on the following criteria:

- Do the articles discuss trust, specifically trust measurement methodologies, or trust-building models in the context of social robotics?
- Does the research address trust repair and its effects on users? This information is crucial for developing effective design guidelines. It enables

us to identify the most effective trust repair strategies for specific situations and integrate them into our design guidelines.
- Only articles on social or service robots were included. All other areas of robotics were excluded (e.g., surgical, industrial, rescue, autonomous vehicles, drones, etc.).
- Each paper must be in English.

3 Trust

Research on trust spans diverse fields including social sciences, psychology, and engineering, with the aim of understanding human-to-human and human-to-machine relationships. In psychology and sociology, the focus of trust research is more on human aspects like gender, age, race, education, expectancy, self-efficacy, and competence. In contrast, for computer science and engineering systems to be trusted, they must be demonstrated safe, reliable, and secure against unauthorized access. Intelligent robots may need to show more indicators of safe behavior around humans [18, 19]. This section will provide a comprehensive scoping review of trust in robotics. This review of trust has been divided into three main parts such as trust building, trust measurement, and trust repair. Additionally, we will explore relevant literature connected to the domains of social science and psychology considering robotics applications.

3.1 Trust Building

In the context of HRI, trust influences the way humans and robots engage and work together. However, establishing trust between users and robots involves more than ensuring the robot is technically reliable or trustworthy. The process of building trust can be divided into two concepts: initial engagement or pre-established trust, and dynamic trust [20, 21]. Pre-established trust is developed prior to direct interaction with the robot. This trust is established based on information that the user has learned or gathered about the system supporting the robot. It implies that users can form initial opinions and expectations about the robot's reliability, capabilities, or other aspects before engaging with it. This prior knowledge or system-related information serves as the foundation for the trust that users develop in anticipation of their interaction with the robot. On the contrary, dynamic trust is based on continuous feedback during real time HRI. It is shaped by past experiences and adjusts in response to various decisions and events occurring throughout the interaction with the robot [21, 22].

A robot may be designed to explain its actions in various scenarios and the impact of such explanations on trust in human-robot interactions is substantial. They are instrumental in restoring trust when a robot’s actions do not meet human expectations, and the nature of the explanation can affect the extent of trust recovery [23]. The context and the type of explanation offered can significantly impact the level of trust generated. For instance, [24] found when opening a bottle that joint visualizations emerge as the most effective form of explanation for enhancing human trust, while text summaries were least effective in that context. Robots usually depend on established rules or

machine learning techniques to identify errors in their functioning. They might also employ sensors and feedback mechanisms to track their performance and pinpoint issues. Nonetheless, robots may not always detect errors on their own and might require human input or additional information to do so. In HRIs, explanations become necessary when the actions of the robot lack clarity or have a considerable effect on the user. The amount of explanation required can differ such as in certain comprehensive technical details can help, while in other cases robots may need to provide straightforward and high-level information. The objective is to provide sufficient information to build trust and comprehension without causing frustration [25]. The appearance of a robot, its movements during interaction (including actions like touching the user [26]), explainability [19], transparency [27], reliability [28], safety [29], accountability [30], the distinction between interactive and non-interactive robots [31, 32], and factors like gender and humor [33] all influence user trust during HRI.

Researchers also investigated numerous models for measuring trust, which we will discuss in the next section. Before addressing trust measurement approaches, it is important to understand the difference between trust building and trust measurement models.[21][6]. Trust building models aim to create and enhance trust, while trust measurement models aim to assess and quantify trust. Trust building models focus on designing interactions and behaviors that create trust, whereas trust measurement models focus on evaluating trust through various metrics and observations.

3.2 Trust Measurement

The study of trust in robotics raises several important questions. It is challenging to incorporate the abstract concept of trust into the design process, especially for researchers who are not experienced in qualitative and quantitative (e.g. surveys) research methods. Additionally, trust is a complex concept with multiple dimensions, wherein the numbers of both trustors (users) and trustees (robots) can vary [18]. To measure and understand trust, researchers often collect subjective data from human participants who interact with robots using various methods, such as interviews, questionnaires, or observations. Additionally, they also utilize methods like Markov processes, probabilistic models, and machine learning prediction models to forecast trust [5]. This data collected from subjective approach is then used to build trust models that can capture and predict the level of trust in different situations and contexts.

Trust measurement and modelling has been approached in a variety of ways. A personalized trust prediction model based on Bayesian inference was proposed by [14]. This trust model allows for continuous updating of trust levels as new information becomes available. In this model, each human participant starts with an initial level of trust in the robot, which can be influenced by prior experiences or preconceived notions. As humans interact with the robot, they observe its performance (including task accuracy, reliability, and efficiency). Using Bayesian inference, the model updates the trust level based on the observed performance. The model is personalized for each user, meaning it considers individual differences in how trust is developed and adjusted. This personalization is achieved by learning the parameters specific to each user's trust dynamics. The multidimensional measure of trust (MDMT) uses a different approach

[31]. It is an intuitive measure of trust that assesses different dimensions of trust in agents in human-robot and human-human trust situations. It includes subscales for reliability, competence, ethics, transparency, and benevolence, organized into broader factors of performance and moral trust. The input size for the MDMT is 20 items. Each item is rated on an 8-point discrete scale, ranging from 0 (Not at all) to 7 (Very), and the output is a set of scores across above mentioned dimensions.

In subjective methods, certain questionnaire-based approaches are tied to well established models, whereas others lack this association. For example, in [34] researchers used a questionnaire to assess the level of trust participants placed in a robot during a collaborative task where participants had to sort laundry with robot assistance. The researchers analyzed the questionnaire data employing various statistical techniques (T-tests, ANOVA, and linear regression) to compare groups, to understand the relationship between trust levels and other continuous variables, and to identify differences in trust levels among multiple groups. Questionnaires associated with trust models are also commonly used. For example, [28] evaluated trust in a real-world mobile navigation scenario involving the utilization of an autonomous wheelchair for delivering packages to predefined locations. The assessment incorporated several models, including the Trust in Automated Systems Test (TOAST) [35], Trust in Automation (TiA) [36], and Trust Perception Scale (TPS) [37][28]. TOAST is a nine-item scale with two main categories: understanding and performance. TIA is a questionnaire developed to measure trust in automation, based on a theoretical model containing six underlying dimensions. TPS is a scale consisting of 40 items and provides a percentage trust score on a 0-100% rating scale.

There is no definitive answer to which trust model is the most used or the best one, as different models may have different advantages and limitations depending on the research question, the application domain, and the type of robot. The personalized trust prediction model is good at capturing trust dynamics over time, but it may not generalize well across different types of HRI. This model also relies on Bayesian inference which can make it complex to understand and apply. It also requires a substantial amount of interaction data to predict trust accurately which could be difficult to get in some cases. MDMT is not able to capture the dynamic nature of trust that evolves over time with continues interaction. It can only measure the trust as a snapshot of the interaction which is not a feasible option in social and service robot as these robots interact with users over the long period of time. The combination of TOAST, TIA, and TPS has its own limitations such as risk assessment and specificity. Trust measured by these scales may not fully account for varying levels of risk associated with different automated systems. While questionnaires can be administered at several time steps, they do not necessarily capture dynamic trust. Simply measuring trust at different intervals does not capture the adaptive process of dynamic trust unless the data is used to adjust behaviors or interactions in real time.

Through a thorough examination of existing literature, a consistent finding emerges: the measurement of trust is not merely a mathematical operation and there is not one model which can fit all trust measurements. The choice of a trust model should be based on the specific requirements of the situation and the type of trust relationship being assessed. Measuring trust effectively requires the participation of human participants.

However, it is important to recognize that involving humans in trust measurement processes adds an additional delay, mainly because approvals from social and ethical committees take time. This extra step, while crucial for ethical considerations, adds a time constraint to research efforts.

3.3 Failure and Trust Repair

Failures are part of HRI, and they can damage the trust that humans have in technology. Different types of failures require different types of trust repair strategies, and different methods are used to evaluate the effectiveness of these strategies. In this section, we focus on the effects of failure types on trust repairs in HRIs and examine how failures and repair methods impact user trust and blame attributions during these interactions. Logic failures are identified as the most critical type of performance failure, significantly impacting the trust between humans and robots [38]. The research suggests that when robots fail to perform tasks as expected due to a reasoning or decision-making error, it severely damages user trust. To address these failures, the studies suggest that an internal attribution apology is the most effective strategy for trust repair. In internal attribution apology, a robot takes responsibility for the mistake. This approach not only enhances user's trust in the robot's competence and integrity but also decreases the perceived severity of the trust breach [39]. This finding is particularly noteworthy as it contrasts with the established norms in human-human and human-machine trust repair, where denial of fault was previously assumed to be more effective.

Another theme discussed in selected literature is trust loss due to deception. The study of deception in robotics is crucial because it addresses the potential consequences of robots behaving in ways that are unexpected or misleading to humans. This is particularly relevant in real-world applications where trust is a foundational aspect of HRI. For instance, in the context of physical rehabilitation, scientists designed a robot that intentionally misled participants about their effort levels, motivating them to work harder and improve their overall rehabilitation outcomes [38, 39]. The intentional error can lead to a mismatch between the robot's actions and its intentions, which is critical to address because it can reduce trust which is a key component in the effectiveness of human-robot collaborations. To navigate these challenges, researchers have proposed a framework for repairing trust that categorizes strategies as either instrumental, which focuses on the outcome of the deception, or relational, which concentrates on the relationship between the human and the robot. Additionally, these strategies are further classified based on whether they align with or contradict the type of deception involved. Empirical studies have been conducted to explore these concepts. For instance, an experiment involving a humanoid robot that deceives participants in a trivia game provided insights into how trust can be rebuilt following deception. They measured trust recovery through self-reported questionnaires and perception scales [40].

Another study investigated the effects of robot deception on human trust and the effectiveness of various apology strategies to repair that trust. This study focuses on a high-stakes, time-sensitive assisted driving scenario where participants interact with a robotic assistant that provides potentially deceptive advice. Participant's trust in the robotic assistant was significantly affected when they were deceived. Different text-

based apologies were tested to see which were most effective at repairing trust. An apology that did not acknowledge intentional deception was found to be the best at mitigating negative influences on trust [41]. This research adds valuable information to the understudied area of robot deception and could guide designers and policymakers in developing artificial intelligence (AI) systems that interact with humans, particularly in situations where trust is crucial. As we discussed in the section 3.1, a robot can offer explanations for their actions. These explanations, whether they focus on the functionality or the mechanics of the task, can be tailored to the users' needs and preferences, thereby aiding in the restoration of trust after a failure [24].

The main insights from this literature are that failures are part of HRI and can have negative impacts on human trust. However, there are still many open questions and challenges in this field, such as how to design and implement effective and ethical trust repair strategies, how to build and measure trust dynamics in HRI, and how to account for individual and cultural differences in human responses to failures and repairs.

4 Design Principles

To design a robot that can be trusted, one must consider the varying degrees of trust needed for different applications. We acknowledge that trust is a complex and dynamic phenomenon that depends on various factors, such as the application domain, the user profile, and the interaction context. Based on our scoping review, we derived the following design principles that could help robots to achieve user's trust.

4.1 Ensure User's Data Privacy

Robots should only collect data that is essential for their function, avoiding any unnecessary data that could violate user privacy. Studies have shown that trust is significantly impacted by how robots handle user data. For instance, research by [42–44] emphasizes that robots should only collect data essential for their function, avoiding any unnecessary data that could violate user privacy. These studies collectively highlight that minimizing data collection to only what is necessary can enhance user trust. Furthermore, robot providers should provide users with simple mechanisms to view, manage, and delete their data, ensuring they have control over their personal information [43, 45, 46]. This mechanism could be a log file which could also be helpful for accountability (see section 4.5).

4.2 Security in Design

After collecting the required data, the next aim of the robot should be to secure it in a safe place to become a trustworthy system. This storage system should not delay the overall quality of service (QoS) of robot. It should have a robust security measure mechanism to protect user data from unauthorized access, breaches, and leaks. Moreover, it is also important to inform your user about potential security threats and risks at the

start. There should be an automatic system to scan the system for any malware in the non-active hours of the robot [43, 45, 46]. For example, a healthcare robot that collects patient data during interactions. After collecting this sensitive information, the robot immediately encrypts the data and stores it in a secure storage system (e.g. cloud). The service providers should ensure that the data is protected from any unauthorized access and breaches.

4.3 Perceived Safety and Actual Safety

In HRI literature, trust emerges as a critical factor influencing perceived safety in interactions between humans and robots. Trust is closely intertwined with other key factors, such as the context of robot use, user comfort, experience and familiarity with robots, a sense of control over the interaction, and transparent and predictable robot actions. These factors are not only essential for ensuring safety but also for building and maintaining user trust, which directly impacts how safe users feel when interacting with robots [29]. Additionally, studies have highlighted the need for redefining safety in light of human-robot interaction, considering not only physical risks but also psychosocial and cybersecurity aspects [47]. In terms of physical safety, the literature has emphasized the importance of service robots providing physically safe services, especially in the context of the COVID-19 pandemic and has proposed a typology of safety-related robot roles [48]. Another study has discussed the ethical design of social robots in aged care, proposing design principles that consider both the physical and emotional well-being of users [49]. Therefore, researchers should develop robots with robust safety protocols for accident prevention, capable of safe human and object detection, and designed with user-trusted features for reassurance and respecting personal space.

4.4 Transparency and Explainability

Transparent communication is key to improving the user's trust in robots [40, 41]. For social and service robotics, the transparency of a robot's actions is important to ensure safety and efficiency because of the shared environment. It is essential for robots to possess planning systems that actively articulate their decision-making process to the user so the user can predict future action [40]. A trustworthy robot should communicate proactively whenever its actions have direct implications for the users or when a decision requires user input or consent. It should also communicate reactively in response to user queries or changes in the environment that affect its operation [50]. This level of openness not only builds trust but also clarifies the robot's intentions, making its behavior predictable and understandable. To enhance transparency, robots must offer clear and logical insights into their actions and strategies. For technical users, detailed explanations of the robot's decision-making process may be appropriate. For general users, the robot should provide simplified, understandable explanations that convey the rationale behind actions without overwhelming the user with technical details. Where possible, the robot should support multimodal communication, consisting of verbal, visual, and possibly tactile signals to convey information effectively (see section 4.7).

4.5 Accountability

Research has shown that a robot's ability to justify its actions and decisions significantly impacts user trust. In [30, 51] researchers emphasize the importance of robots maintaining secure logs of their operations and events, ensuring data protection and retrievability. When robots communicate their activities through effective channels like verbal communication, body language, or visual signals, users can better understand and trust their actions. They found that transparency in a robot's decision-making process significantly increases user trust and safety in shared environments. The studies concluded that proactive communication from robots about actions that directly impact users or require user input enhances trust. Additionally, the robot's accountability mechanisms should operate without significantly affecting its efficiency or consuming excessive resources. For example, a social companion robot that interacts with individuals in environments such as residences, medical facilities, and eldercare centers. To adhere to accountability standards, the robot incorporates several functionalities that enable it to articulate and document its actions. For instance, the robot can audibly rationalize its choices, such as reminding a patient to take their medication at a designated time. It employs a decision-tracking system that logs the thought process behind each decision, accessible to those with proper authorization. Furthermore, the robot keeps a protected electronic journal of all its interactions and occurrences. This journal is encrypted and stored in compliance with privacy laws and data protection standards.

4.6 Apology Mechanism

The robot should have a built-in mechanism to generate an apology after detecting a failure. The apology should be appropriate to the context of the error and the impact it may have had on the user. The apology should also be consistent with what the user would normally expect in a similar situation involving human-to-human interaction. The robot's acknowledgment of fault should be in line with the user's attributions, meaning that the robot should not only apologize for the error but also provide an explanation where possible that matches the user's understanding of why the error occurred [38, 51, 52]. Furthermore, in scenarios where the robot does not detect its own failure, it should be capable of acknowledging and responding to user-identified errors. Upon receiving such feedback, the robot should offer an apology that aligns with the user's understanding of the situation. For instance, if a home assistant robot fails to execute a task correctly or causes an accident like spilling a beverage, it should promptly admit the error and activate its apology sequence.

4.7 Communication and Interaction

The robot should communicate with the human user in a clear, timely, and respectful way. The robot should use verbal and non-verbal cues, such as speech, gestures, and facial expressions, to convey its messages and events [45, 53]. The robot should also respect the user's privacy and autonomy and avoid interrupting or disturbing the user unnecessarily. Take, for example, a robot working in a care setting. In a quiet early

morning environment, it could acknowledge residents with a nod and a softly spoken "Good morning". It may use encouraging body language, such as open hands to denote a readiness to assist, and it can establish simulated eye contact through its screen to show engagement.

Cultural contexts are also crucial in the robot's design. In environments where direct eye contact is seen as hostile, the robot's visual interaction must be modified. The robot's speech patterns, and accent should also be customized to align with local dialects where possible to enhance understanding. Moreover, robot gestures should also be designed to fit in the cultural norms to prevent any potential offense; for example, a thumbs-up gesture may be positive in some cultures but not in others. By accommodating these cultural differences, robots offer a more respectful interaction, thereby improving the human-robot relationship.

4.8 Utility and Ease of Use

The robot should be perceived as useful and easy to use by the user and provide value and support for the user or the task. These factors are important for fostering trust in robots, especially for social assistive robots (SARs) that aim to help and interact with humans [52, 54]. Consider in a rehabilitation center, a robot is introduced to assist patients with their recovery exercises. It is designed with a user-friendly interface that patients can interact with using touch or voice commands. It guides them through their exercises, provides feedback on their progress, and adjusts routines according to their recovery rate, making it an invaluable aid for both patients and therapists. It also features a friendly avatar that encourages patients with positive affirmations and celebrates their milestones, fostering a sense of companionship. By providing consistent support, understanding individual needs, and engaging in a socially meaningful way, it builds trust with patients, which is essential for the success of social assistive robots in healthcare settings.

4.9 Visual Appearance

The design of a robot is largely influenced by its intended use. For instance, if the goal is to develop an assistive robot, an anthropomorphic design, which resembles human characteristics, is often the preferred choice [55, 56]. On the other hand, if the aim is to create a pet or therapeutic robot, a zoomorphic design, which mimics animal characteristics, may be more suitable. When it comes to social and service robots, those designed with human-like features may be more easily accepted by the user. This is because human-like characteristics can trigger social behaviors and responses from people, making the interaction with the robot feel more natural and intuitive. When a robot has a face, eyes, or can mimic human gestures, it helps to bridge the gap between mechanical and social entities, allowing users to connect with the robot in a way that is like human-to-human interaction [57, 58]. For example, the robot Pepper is designed to assist humans in various tasks, such as providing information and interacting socially. Its anthropomorphic design, with a human-like face and body, helps it to be more relatable and approachable for users [59].

Although this design principle is important in social and service robotics, the existing research literature does not provide sufficient evidence or guidance to determine the specifications for a new robot design.

5 Conclusion

In conclusion, this paper has conducted a comprehensive interdisciplinary scoping review of research on human trust models, measurement of trust, as well as failure and repair in robotics. Based on the scoping review we identify nine principles that could aid the development of trustworthy service and social robots. Our research indicates that there is no single trust model that fits in all scenarios. Different models have their own unique strengths and weaknesses, which can be influenced by various factors such as the research question, the application domain, and the type of robot. The literature highlights that failures are an integral part of HRI and can negatively impact human trust and behavior. Furthermore, different types of failures call for specific repair strategies, the success of which depends on several factors including the nature and severity of the failure, the context, task requirements, and the characteristics of both humans and robots. However, there are still many unanswered questions and challenges in this field, such as the design and implementation of effective trust repair strategies, the measurement and modeling of trust dynamics in HRI, and the understanding of individual and cultural differences in human responses to failures and repairs.

Acknowledgements. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 18/CRT/6222.

References

1. Hupcey, J.E., Penrod, J., Morse, J.M., Mitcham, C.: An exploration and advancement of the concept of trust. J Adv Nurs. 36, 282–293 (2001).
2. Mejia, C., Kajikawa, Y.: Bibliometric analysis of social robotics research: identifying research trends and knowledgebase. Applied Sciences. 7, 1316 (2017).
3. Henschel, A., Laban, G., Cross, E.S.: What makes a robot social? a review of social robots from science fiction to a home or hospital near you. Current Robotics Reports. 2, 9–19 (2021).
4. Hancock, P.A., Kessler, T.T., Kaplan, A.D., Brill, J.C., Szalma, J.L.: Evolving trust in robots: specification through sequential and comparative meta-analyses. Hum Factors. 63, 1196–1229 (2021).
5. Kok, B.C., Soh, H.: Trust in robots: Challenges and opportunities. Current Robotics Reports. 1, 297–309 (2020).
6. Kraus, J., Miller, L., Klumpp, M., Babel, F., Scholz, D., Merger, J., Baumann, M.: On the role of beliefs and trust for the intention to use service robots: an integrated trustworthiness beliefs model for robot acceptance. Int J Soc Robot. 1–24 (2023).
7. Naneva, S., Sarda Gou, M., Webb, T.L., Prescott, T.J.: A systematic review of attitudes, anxiety, acceptance, and trust towards social robots. Int J Soc Robot. 12, 1179–1201 (2020).

8. Schaefer, K.: The perception and measurement of human-robot trust. (2013).
9. Robinette, P., Howard, A.M., Wagner, A.R.: Timing is key for robot trust repair. In: Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7. pp. 574–583. Springer (2015).
10. 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).
11. Stower, R., Calvo-Barajas, N., Castellano, G., Kappas, A.: A meta-analysis on children's trust in social robots. Int J Soc Robot. 13, 1979–2001 (2021).
12. Brzowski, M., Nathan-Roberts, D.: Trust measurement in human–automation interaction: A systematic review. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. pp. 1595–1599. SAGE Publications Sage CA: Los Ange-les, CA (2019).
13. Hancock, P.A., Kessler, T.T., Kaplan, A.D., Stowers, K., Brill, J.C., Billings, D.R., Schaefer, K.E., Szalma, J.L.: How and why humans trust: A meta-analysis and elaborated model. Front Psychol. 14, (2023).
14. Guo, Y., Yang, X.J.: Modeling and predicting trust dynamics in human–robot teaming: A Bayesian inference approach. Int J Soc Robot. 13, 1899–1909 (2021).
15. Vorm, E.S., Combs, D.J.Y.: Integrating transparency, trust, and acceptance: The intelligent systems technology acceptance model (ISTAM). Int J Hum Comput Interact. 38, 1828–1845 (2022).
16. Munn, Z., Peters, M.D.J., Stern, C., Tufanaru, C., McArthur, A., Aromataris, E.: Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. 18, 1–7 (2018).
17. Arksey, H., O'Malley, L.: Scoping studies: towards a methodological framework. Int J Soc Res Methodol. 8, 19–32 (2005).
18. Krausman, A., Neubauer, C., Forster, D., Lakhmani, S., Baker, A.L., Fitzhugh, S.M., Gremillion, G., Wright, J.L., Metcalfe, J.S., Schaefer, K.E.: Trust measurement in human autonomy teams: Development of a conceptual toolkit. ACM Transactions on Human-Robot Interaction (THRI). 11, 1–58 (2022).
19. Emaminejad, N., Akhavian, R.: Trustworthy AI and robotics: Implications for the AEC industry. Autom Constr. 139, 104298 (2022).
20. de Pagter, J.: From EU Robotics and AI governance to HRI Research: implementing the Ethics Narrative. Int J Soc Robot. 1–15 (2023).
21. Zhang, W., Wong, W., Findlay, M.: Trust and robotics: a multi-staged decision-making approach to robots in community. AI Soc. 1–16 (2023).
22. Miller, L., Kraus, J., Babel, F., Baumann, M.: More than a feeling interrelation of trust layers in human-robot interaction and the role of user dispositions and state anxiety. Front Psychol. 12, 592711 (2021).
23. Bai, Z., Chen, K.: Effects of Explanations by Robots on Trust Repair in Human-Robot Collaborations. In: International Conference on Human-Computer Interaction. pp. 3–14. Springer (2024).
24. Edmonds, M., Gao, F., Liu, H., Xie, X., Qi, S., Rothrock, B., Zhu, Y., Wu, Y.N., Lu, H., Zhu, S.-C.: A tale of two explanations: Enhancing human trust by explaining robot behavior. Sci Robot. 4, eaay4663 (2019).
25. Stange, S., Hassan, T., Schröder, F., Konkol, J., Kopp, S.: Self-explaining social robots: an explainable behavior generation architecture for human-robot interaction. Front Artif Intell. 5, 866920 (2022).

26. Law, T., Malle, B.F., Scheutz, M.: A touching connection: how observing robotic touch can affect human trust in a robot. Int J Soc Robot. 1–17 (2021).
27. Zerilli, J., Bhatt, U., Weller, A.: How transparency modulates trust in artificial intelligence. Patterns, 100455, (2022).
28. Lingg, N., Demiris, Y.: Building trust in assistive robotics: Insights from a real-world mobile navigation experiment. In: Proceedings of the First International Symposium on Trustworthy Autonomous Systems. pp. 1–7 (2023).
29. Akalin, N., Kiselev, A., Kristoffersson, A., Loutfi, A.: A taxonomy of factors influencing perceived safety in human–robot interaction. Int J Soc Robot. 15, 1993–2004 (2023).
30. Fernández-Becerra, L., Guerrero-Higueras, Á.M., Rodríguez-Lera, F.J., Fernández-Llamas, C.: Analysis of the performance of different accountability strategies for autonomous robots. In: 14th International Conference on Computational Intelligence in Security for Information Systems and 12th International Conference on European Transnational Educational (CISIS 2021 and ICEUTE 2021) 14. pp. 41–51. Springer (2022).
31. Graf, L., Torkar, M., Stückelmaier, E., Sichler, R., Malafosse, P., Fischer, K., Palin-ko, O.: Perceived Trustworthiness of an Interactive Robotic System. In: 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 773–777. IEEE (2022).
32. Gurung, N., Herath, D., Grant, J.B.: Feeling safe: A study on trust with an interactive robotic art installation. In: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. pp. 447–451 (2021).
33. Smakman, M.H.J., Vanegas, D.F.P., Smit, K., Leewis, S., Okkerse, Y., Obbes, J., Uffing, T., Soliman, M., van der Krogt, T., Tönjes, L.: A trustworthy robot buddy for primary school children. Multimodal Technologies and Interaction. 6, 29 (2022).
34. Gideoni, R., Honig, S., Oron-Gilad, T.: Is it personal? The impact of personally relevant robotic failures (PeRFs) on humans' trust, likeability, and willingness to use the robot. Int J Soc Robot. 1–19 (2022).
35. Wojton, H.M., Porter, D., T. Lane, S., Bieber, C., Madhavan, P.: Initial validation of the trust of automated systems test (TOAST). J Soc Psychol. 160, 735–750 (2020).
36. Körber, M.: Theoretical considerations and development of a questionnaire to measure trust in automation. In: Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20. pp. 13–30. Springer (2019).
37. Schaefer, K.E.: Measuring trust in human robot interactions: Development of the "trust perception scale-HRI." In: Robust intelligence and trust in autonomous systems. pp. 191–218. Springer (2016).
38. Zhang, X., Lee, S.K., Maeng, H., Hahn, S.: Effects of Failure Types on Trust Repairs in Human–Robot Interactions. Int J Soc Robot. 15, 1619–1635 (2023).
39. Brewer, B.R., Fagan, M., Klatzky, R.L., Matsuoka, Y.: Perceptual limits for a robotic rehabilitation environment using visual feedback distortion. IEEE transactions on neural systems and rehabilitation engineering. 13, 1–11 (2005).
40. Rosero, A.: Using Justifications to Mitigate Loss in Human Trust when Robots Perform Norm-Violating and Deceptive Behaviors. In: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. pp. 766–768 (2023).
41. Rogers, K., Webber, R.J.A., Howard, A.: Lying About Lying: Examining Trust Repair Strategies After Robot Deception in a High-Stakes HRI Scenario. In: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. pp. 706–710 (2023).
42. Chatzimichali, A., Harrison, R., Chrysostomou, D.: Toward privacy-sensitive hu-man–robot interaction: Privacy terms and human–data interaction in the personal robot era. Paladyn. 12, 160–174 (2020).

43. Lutz, C., Tamò-Larrieux, A.: Do privacy concerns about social robots affect use intentions? Evidence from an experimental vignette study. Front Robot AI. 8, 627958 (2021).
44. Akalin, N., Kristoffersson, A., Loutfi, A.: Evaluating the sense of safety and security in human–robot interaction with older people. Social robots: Technological, societal and ethical aspects of human-robot interaction. 237–264 (2019).
45. Fronemann, N., Pollmann, K., Loh, W.: Should my robot know what's best for me? Human–robot interaction between user experience and ethical design. AI Soc. 37, 517–533 (2022).
46. Yang, D., Chae, Y.-J., Kim, D., Lim, Y., Kim, D.H., Kim, C., Park, S.-K., Nam, C.: Effects of social behaviors of robots in privacy-sensitive situations. Int J Soc Ro-bot. 1–14 (2022).
47. Martinetti, A., Chemweno, P.K., Nizamis, K., Fosch-Villaronga, E.: Redefining safety in light of human-robot interaction: A critical review of current standards and regulations. Frontiers in chemical engineering. 3, 666237 (2021).
48. Schepers, J., Streukens, S.: To serve and protect: a typology of service robots and their role in physically safe services. Journal of Service Management. 33, 197–209 (2022).
49. Yuan, S., Coghlan, S., Lederman, R., Waycott, J.: Ethical Design of Social Robots in Aged Care: A Literature Review Using an Ethics of Care Perspective. Int J Soc Robot. 15, 1637–1654 (2023).
50. Che, Y., Okamura, A.M., Sadigh, D.: Efficient and trustworthy social navigation via explicit and implicit robot–human communication. IEEE Transactions on Robotics. 36, 692–707 (2020).
51. Fratczak, P., Goh, Y.M., Kinnell, P., Justham, L., Soltoggio, A.: Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction. Int J Ind Ergon. 82, 103078 (2021).
52. Zafrani, O., Nimrod, G., Edan, Y.: Between fear and trust: Older adults' evaluation of socially assistive robots. Int J Hum Comput Stud. 171, 102981 (2023).
53. Kim, T., Song, H.: "I Believe AI Can Learn from the Error. Or Can It Not?": The Effects of Implicit Theories on Trust Repair of the Intelligent Agent. Int J Soc Ro-bot. 15, 115–128 (2023).
54. Schwaninger, I., Güldenpfennig, F., Weiss, A., Fitzpatrick, G.: What do you mean by trust? Establishing shared meaning in interdisciplinary design for assistive technology. Int J Soc Robot. 13, 1879–1897 (2021).
55. Elshan, E., Zierau, N., Engel, C., Janson, A., Leimeister, J.M.: Understanding the design elements affecting user acceptance of intelligent agents: Past, present and future. Information Systems Frontiers. 24, 699–730 (2022).
56. Li, Y., Zhou, X., Jiang, X., Fan, F., Song, B.: How service robots' human-like appearance impacts consumer trust: a study across diverse cultures and service settings. International Journal of Contemporary Hospitality Management. (2024).
57. Rossi, A., Holthaus, P., Perugia, G., Moros, S., Scheunemann, M.: Trust, acceptance and social cues in human–robot interaction (SCRITA), (2021).
58. Esterwood, C., Essenmacher, K., Yang, H., Zeng, F., Robert, L.P.: A personable robot: meta-analysis of robot personality and human acceptance. IEEE Robot Au-tom Lett. 7, 6918–6925 (2022).
59. Doncieux, S., Chatila, R., Straube, S., Kirchner, F.: Human-centered AI and robotics. AI Perspectives. 4, 1 (2022).