UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

"SORRY, IT WAS MY FAULT": REPAIRING TRUST IN HUMAN-ROBOT INTERACTIONS

A THESIS SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

MASTER OF ARTS

By XINYI ZHANG

Norman, Oklahoma

2021

“SORRY, IT WAS MY FAULT”: REPAIRING TRUST IN HUMAN-ROBOT INTERACTIONS

A THESIS APPROVED FOR THE

DEPARTMENT OF COMMUNICATION

BY THE COMMITTEE CONSISTING OF

Dr. Sun Kyong Lee, Chair

Dr. Claude Miller

Dr. Amy Johnson

Contents

Abstract

Robots have been playing an increasingly important role in human life, but their performance is yet far from perfection. Based on extant literature in interpersonal, organizational, and human-machine communication, the current study develops a three-fold categorization of technical failures (i.e., logic, semantic, and syntax failures) commonly observed in human-robot interactions from the interactants' end, investigating it together with four trust repair strategies: internal-attribution apology, external-attribution apology, denial, and no repair. The 743 observations conducted through an online experiment reveals there exist some nuances in participants' perceived division between competence- and integrity-based trust violations, given the ontological differences between humans and machines. The findings also suggest prior propositions about trust repair from the perspective of attribution theory only explain part of the variance, in addition to some significant main effects of failure types and repair methods on HRI-based trust.

Keywords: human-robot interactions, technical failures, trust repair, blame attribution

"Sorry, It Was My Fault": Repairing Trust in Human-Robot Interactions

As technology becomes more deeply involved in human life, the relationships between humans and technology have grown more interdependent (Guzman & Lewis, 2020). Consequently, trust is no longer a socio-psychological concept only applicable to interpersonal dynamics. Akin to trust developed through human-to-human communication, trust toward technology also reflects trustors' evaluations of trustees' abilities and helpfulness in achieving expected goals. Since trust is closely associated with interaction outcomes and usage decisions (Sanders et al., 2019)—misgauged levels of trust in technology might lead to misuse, disuse, and abuse of technological systems, while accurately calibrated trust can assist human-machine collaborations (Parasuraman & Riley, 1997)—trust evolved in human-machine communication (HMC), including human-automation interactions, human-agent interactions, human-computer interactions, and human-robot interactions (HRI), has attracted significant scholarly interest.

The present study specifically focuses on trust in robotics, which presents a relatively novel scope in the discipline compared to trust research in automation (Schaefer, 2013; Baker et al., 2018). Following Lee and See (2004), the current study defines trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p. 54). Distinct from interpersonal trust, trust in robots is characterized by unique expectations with a heavy emphasis on system performance (Baker et al., 2018; Hancock et al., 2011) and situational risks and uncertainty (Schaefer, 2013). Previous studies have revealed that performance is the central predictor of human trust in robots (Hancock et al., 2011). Existing robotic performance, however, can hardly reach perfection since numerous errors can occur in human-robot interactions, such as failing to provide responses, mistaking voice commands, identifying physical surroundings inaccurately, and producing incorrect output, leading to decline of trust (Desai et al., 2012; Desai et al., 2013; Salem et al., 2015). Given these issues, it becomes particularly important for roboticists to enhance robots' capabilities for detecting performance failures and repairing human users' trust in order to facilitate effective human-robot interactions (Brooks, 2017; Sebo et al., 2019).

Humans employ various strategies to rehabilitate interpersonal trust, including apologizing, denying, promising, emphasizing, and explaining, and these strategies could potentially be transplanted to the HRI context (de Visser et al., 2018). Essentially, these strategies repair trust by redirecting attributions of blame (Tomlinson & Myer, 2009) and mitigating negative influences of expectancy violations (Lee et al., 2010). In organizational literature, Kim et al. (2004) identified two types of trust violations, competence- and integrity-based violations (details discussed below), and investigated appropriateness of two repair methods, apology and denial, respectively under each condition. The study identified apology as the optimum response for competence-based violations whereas denial is more effective with integrity-based violations, and Sebo et al.'s (2019) study also confirmed this finding in HRI. Yet due to ontological differences between humans and robots, this study questioned whether interactants will perceive robotic errors in the same way as they perceive human errors, since robots are thought of as mindless beings with less agency (Banks, 2019; Gray et al., 2012).

The current study identified three types of technical failures resulting from basic system errors (i.e., logic, semantic and syntax errors; McCall & Kölling, 2014) after revising the human-automation trust repair framework proposed by Marinaccio et al. (2015). Next, the present study explored effectiveness of three trust recovery tactics, apology with internal or external attributions and denial, taking no repair as a reference point, and investigated the potential interaction between failure types and research methods.

Trust Violations and Repairs

Although trust formation has conventionally been understood as a long-term process, studies have disclosed that trust can also develop within a short period of time in temporary groups (Meyerson et al., 1996; Robert et al., 2009). As a critical factor for development and maintenance of social and professional relationships (Haesevoets et al., 2015), trust has long been a topic of intense research across various disciplines (Lewicki & Brinsfield, 2017): by 2013, there had been over 300 documented definitions of trust (Schaefer, 2013). These definitions have conceptualized trust as beliefs, attitudes, intensions, and behaviors, and such a variety of views can be eventually reconciled since attitude is the elemental base of all other dimensions of trust (Lee & See, 2004). According to the classic model presented by Mayer et al. (1995), this perceptual construct is essentially a function of three characteristics of trustees: ability (i.e., “that group of skills, competencies, and characteristics that enable the party to have influences within some specific domain”, p.717), benevolence (i.e., “the extent to which a trustee is believed to want to do good for the trustor”, p.718), and integrity (“the trustor’s perception that the trustee adheres to a set of principles that the trustor finds acceptable”, p.719).

Even though humans and machines are much different entities, some parallels exist between trust fostered in human-to-human communication and trust bred in HMC. For example, many factors contributing to interpersonal trust also influence technology-based trust, such as culture, age, personality for dispositional trust, task difficulty and mood for situational trust, and performance reliability, predictability, error timing, and trustees’ characteristics for learned trust (Hoff & Bashir, 2015). Such similarities mirror the foundational vision of media equation theory and computers-as-social-actor (CASA) paradigm which explain social norms of the human world would be equally applicable in human-to-machine interactions despite ontological differences (Nass & Moon, 2000; Reeves & Nass, 1996).

Development and maintenance of trust is inevitably accompanied by risks of trust violations. Ubiquitous in daily interactions, trust violations can be defined as “unmet expectations concerning another’s behavior, or when the person does not act consistent with one’s values” (Bies & Tripp, p. 248); these violations can vitiate trust, thus resulting in social and economic loss (Rao & Lee, 2007). In response to perceived transgressions, trustees can take attempts to obtain forgiveness (Tomlinson et al., 2004) and restore positive expectations while minimizing negative ones (Kramer & Lewicki, 2010). In contrast to the abundance of trust literature, heretofore research on repair strategies has been a late bloomer, given the common existence of trust violations and subsequent needs for trust repair, though it has gained increased attention in recent years (Dirks et al., 2009). In the field of HMC, trust repair research is still a fledgling subject that seeks theoretical support from interpersonal and organizational trust literature.

Human-to-Human Trust Repair

The current line of trust repair studies in HMC is mainly inspired by a burgeoning line of trust repair research in the field of organizational communication. A series of studies (Ferrin et al., 2007; Kim et al., 2004; 2006; 2013) related to trust repair were initially conducted under a scenario of job interviews, where participants were positioned as a manager looking for a tax accountant and watching the videotapes of a job candidate who performed either competence- or integrity-based violations and attempted to repair trust with different strategies. Apology (i.e., a response in which trustee accepts responsibility, expresses repentance, and stresses the intent to avoid similar violations in the future) and denial (i.e., a response in which the trustee rejects responsibility and expresses no repentance) have been two trust repair methods of primary scholarly interest, and researchers previously obtained mixed findings about their effects (Kim et al., 2004): while some researchers believed that apology repairs trustors' faith more successfully by increasing positivity in perceived intentions and motives, others contended that avoiding the blame would be more effective given the seriousness of accusation (Ferrin et al., 2007; Kim et al., 2006; Takaku, 2001). To reconcile the inconsistencies in prior findings, Kim et al. (2004) introduced two types of violations from a diagnostic perspective: competence-based violations and integrity-based violations.

This two-fold taxonomy of trust violations developed by Kim et al. (2004) rests on the recognition that competence and integrity are two critical determinants of trust. The fundamental differences between two types of violations root in hierarchically restrictive schemas (Reeder & Brewer, 1979). Trustors' evaluations of two types of trust violations follow distinct processes because skill (i.e., competency) and honesty (i.e., integrity) invoke different attribution patterns: competence and performance can be tested, but integrity and morality cannot be easily quantified. In addition, violations of integrity are judged more harshly than the ones against competence because they endanger trustors' comprehensive evaluations of trustees as human beings (Sitkin & Roth, 1993). Since a single good performance is more likely to be regarded a reliable sign of innate competence while one terrible performance might be interpreted as an anomaly caused by situational factors, positive information outweighs negative one for competence-based violations. By contrast, negative information overshadows positive one for integrity-based violations because one honest behavior is not considered to be a dependable indicator of honesty whereas a single dishonest behavior is deemed to be an indicator of dishonesty (Kim et al., 2004). Therefore, the remedies for competence-based violations should focus on maximizing positivity while the remedies for integrity-based violations should focus on minimizing negativity.

Consequently, Kim et al. (2004) assume that apology is more effective with competence-based violations than denial, since trustors pay more attention to positivity (i.e., expressed repentance and the intent to avoid such violations in the future) brought by acknowledgement of such violations rather than negativity (i.e., accepted culpability); in contrast, denial is more potent succor for integrity-based violations, because avoiding negativity (i.e., accepted culpability) regarding such violations would be more effective than generating positivity (i.e., expressed repentance and the intent of redemption). Through their initial studies, Kim et al. (2004) measured trust with two subconstructs, trusting beliefs (i.e., the trustor's perceived competence and integrity of the trustee) and intentions (i.e., the trustor's tendency to rely on the trustee in vulnerability), and identified significant interactions between the two violation types and the two repair methods; they also found both repair methods, especially denial, would backfire when the truth was inconsistent with the claims.

Subsequent studies conducted by Kim and his team mostly substantiated their preliminary findings. Later Kim et al. (2006) incorporated attribution theory into their experimental designs and concluded that apologies with internal attributions produced better outcomes than ones with external attributions for competence-based violations, but the findings were flipped for integrity-based violations. The researchers interpreted those findings to mean that integrity-based violations are so deleterious that any mitigating response, no matter how untenable, will serve as a relief. Ferrin et al. (2007) noted that reticence (i.e., a response in which the trustee claims he or she cannot or will not confirm or deny the responsibility) would be less effective than optimal responses for both violation types in light of the psychological fact that an untested accusation can still dispose people to believe it, verifying hypotheses with two scenarios (i.e., job interviews for a tax accountant and interrogations of an executive officer). Kim et al. (2013) investigated the process of trust recovery in the social context and again observed the same interactions, with group dynamics intervening individual judgment. In the recent decade, trust repair research has proliferated in the field of organizational communication (Bachmann et al., 2015; Eberl et al., 2015; Fuoli et al., 2017; Gillespie & Dietz, 2009; Janowicz-Panjaitan & Krishnan, 2009; Poppo & Schepker, 2010), exploring the issue on multiple organizational levels.

Notable support Kim et al.'s (2004) model has received notwithstanding, a few studies have indicated otherwise: the study results observed by Utz et al. (2009) indicated that a plain apology was considered more believable than a denial for both competence- and integrity-based violations for eBay buyers. Bansal et al. (2015) argued that apology was superior to denial for every type of trust violation (i.e., ability, benevolence, integrity) in a scenario of privacy breach, and denial even performed worse than no-response under certain conditions. Such counterevidence denotes the model established by Kim et al. (2004) might not be equally applicable under certain contexts due to differences in participants' trust patterns since the mechanisms underlying apology and denial are highly complex.

According to Lewicki and Brinsfield (2017), apology, denial, and other alternative methods, including giving verbal accounts, excuses, or explanations and providing tangible compensations, belong to short-term repair strategies, as opposed to long-term ones (e.g., making structural arrangements, reframing violations). Apart from the interactions depicted by Kim et al. (2004), organizational and interpersonal communication scholars have also explored other key elements affecting reconciliation between trustors and trustees. Prior studies have highlighted timing, severity, and frequency of violations (Lewicki & Brinsfield, 2017), dispositional trust

(Colquitt et al., 2007; Kramer, 1999), relationship characteristics (e.g., relative status; Aquino et al., 2001; past relationships and probability of future violations; Tomlinson et al., 2004) to be influential factors in trust repair. Repair tactics such as intensity, perceived sincerity, and multi-dimensionality (i.e., display of regret, explanations, acknowledgement of accountability, offer of future repair, and entreaty for forgiveness) of apology (Lewicki & Tomlinson 2003; Lewicki & Brinsfield, 2017; Tomlinson et al., 2004), as well as timeliness of act, are variables associated with effectiveness of trust repair attempts. Trustees' characteristics also matter, including personal traits, such as likeability (Bradfield & Aquino, 1999) and gender (Walfisch et al., 2013), and organizational features such as pre-crisis reputations (Beldad et al., 2018; Lewicki & Tomlinson, 2003). To summarize, trust repair is a highly complex practice because of the multifaceted nature of trust and contextual variance, and this complexity invites further theoretical and empirical exploration.

Robot-to-Human Trust Repair

Research focusing on trust promotion in HMC is copious, but research into trust in robotic systems is a relatively new emphasis (Baker et al., 2018). Particularly, existing trust literature in HMC mainly sheds light on technical designs (e.g., visual anthropomorphism, machine politeness) that increase baseline trust levels and hence benefit trust resilience (de Visser et al., 2016; Quinn, 2018). Previous research in other areas of HMC can be regarded as a useful starting point for studying HRI-based trust because of the similarities shared amongst technological systems, although robots may possess more advanced capabilities as autonomous entities than conventional automations and virtual agents (de Visser et al., 2018). Interpersonal and organizational communication literature also provides some valuable references based on similarities in trust nurtured by the two types of interactions (i.e., human-to-human and HRI).

In the context of HRI, factors affecting trust development can be roughly classified into human-related (ability-based factors and characteristics), robot-related (performance- and attribute-based factors), and environment-related factors (tasking and team collaboration) according to Hancock and his colleagues (2011). Similar to trust loss in human-to-human communication, trust violations also happen in HRI when robots fail to meet humans' expectations or display mismatched principles and goals (de Visser et al., 2017). Since few, if any, robots attain perfection in their designs (Honig & Oron-Gilad, 2018), and since performance-based factors turn out to be the central determinant in human evaluations of robotics (Hancock et al., 2011), humans' trust in robots is constantly challenged by robotic failures.

Numerous studies indicate that human-to-machine trust declines after machines violate humans' expectations, which is often caused by system failures (Corritore et al., 2003; Desai et al., 2012; 2013; Salem et al., 2015; Sanchez et al., 2014; Vries et al., 2003). Scholars have previously scrutinized the effects of time-, magnitude-, and outcome-based error variation on trust assessment: for instance, Madhavan et al. (2006) found that violations are considered more negatively when tasks are perceived to be easy; Desai et al. (2013) disclosed that reliability drops in the earlier stages of interactions are more harmful than the ones occurring later and predicted that trust inertia (i.e., delayed trust recalibration) also exists in HMC given the discrepancy between real-time and overall trust measures; Rossi et al. (2017) postulated that severity in negative consequences brought by violations determines the magnitude of trust regression.

In response to the prejudicial effects of robotic failures, prior research indicates that robots can initiate trust repair just as humans can. Repair attempts from autonomous systems can also bolster perceived sociability and humanness, further promoting trust resilience (de Visser et al., 2012; 2016; 2018). Such attempts at trust repair may remain effective even when machines' reliability does not really improve (de Visser, 2012), so trust repair is not only effective but is also efficient as far as technology designs are concerned, since machines cannot easily make progress in performance. For trust repair in HMC, previous findings from this line of research mostly adhere to the ones drawn from human-to-human interactions, encompassing various repair strategies, such as ignoring (Correia et al., 2018), blaming (Groom et al., 2010; Kaniarasu & Steinfeld, 2014), apologizing (de Visser et al., 2016; Lee et al., 2010; Quinn, 2018; Robinette et al., 2015; Sebo et al., 2019; Tzeng, 2004; Wagner, 2016), denying (Quinn, 2018; Sebo et al., 2019), promising (Robinette et al., 2015; Wagner, 2016), justifying (Correia et al., 2018), engaging in social dialogues (Lucas et al., 2018), giving palpable compensations (Lee et al., 2010), offering options (Lee et al., 2010) and providing additional information (Robinette et al., 2015) or support (Brooks, 2017).

In general, these studies espouse the perspective that it is better to take repair actions than not, but the investigation has been relatively fragmented (de Visser et al., 2020). With respect to comparison of specific repair strategies, Lee et al. (2010) found that expressions of remorse and promises were more powerful than offers of options after a breakdown in robotic service, with individuals' orientations (relational or utilitarian) determining which strategy was optimal. Wagner (2016) observed that promising for the future was a better booster for human trust in the robot than was apologizing for the past in an emergency excavation task. Besides, researchers have also noted human (e.g., operator attention, age) and contextual factors (e.g., task risks, task difficulty, system reputations, system expertise) significantly influence repair outcomes in this process (Brooks, 2017; Schaefer et al., 2012)

Most importantly, there have been two studies that examine Kim et al.'s (2004) framework: Quinn (2018) found that although apology was more effective as a repair method for competence-based violations than for integrity-based violations, repair outcomes of denial did not differ for two types of violations; Sebo et al. (2019) substantiated the interactions asserted by Kim et al. (2004) with a competitive shooting game in which a robot broke its initial promise and framed the behavior to be either competence- or integrity-based, and the researchers noted that human players were more inclined to retaliate under the condition of integrity-based violations and denial. These findings verify the connections between trust violations and repair methods derived from human-to-human communication, further supporting the symmetry between human-to-human and human-to-robot trust, although the fundamental differences between humans and robots have not been subjected to rigorous consideration.

Uniqueness of HRI-Based Trust

Perceptual Differences

Existing research on trust repair for HMC is deeply rooted in interpersonal and organizational communication literature and has exhibited commonality bridging the two fields. Nevertheless, a few studies have unveiled some key differences between interpersonal trust and HMC-based trust. Humans seem to possess different levels of dispositional trust toward humans and machines, allocating more initial trust to the latter owing to higher perceived authority resulting from bias toward automation (de Visser, 2016; Dzindolet et al., 2003; Parasuraman & Manzey, 2010). For example, Madhavan and Wiegmann (2005) noticed that participants reached more agreement with an automation advisor than with a human even when both were labeled "novice." As a corollary, people might overreact to system failures due to interruptions of perfect automation schema (Dzindolet et al., 2002): in the same study, for instance, Madhavan and

Wiegmann (2005) also found that participants were more likely to notice those errors generated by the system than the ones by humans. Therefore, trust in machines can be harder to reestablish once violated than interpersonal trust (Hoffman et al., 2013) because machine failures can potentially lead to greater negative expectancy violations.

Conceptual differences between humans and technologies also add to perceptual inequivalence toward trust violations and their repairs. First and foremost, the belief that machines are more fixed and less changeable than humans will potentially impair the effects of trust repair, because humans may hold the opinion that an oral repair is not likely to be followed by an actual improvement in performance (de Visser et al., 2018). Particularly, they may perceive trust repair efforts from machines to be less sincere because repair is predefined by algorithms (de Visser et al., 2018), especially when repair attempts appear uniform across different kinds of situations. Second, humans also perceive morality in machines differently because machines do not have human minds, which means they cannot accumulate sensational experiences as humans do (Banks, 2019; Gray et al., 2012); thus, human judgments of machines related to moral principles (i.e., integrity and benevolence) may differ from human judgments of other humans.

Additionally, machines are also perceived to possess less agency and are viewed as less legitimate of making moral decisions (Gray et al., 2007; 2012; Malle et al., 2016). As posited by Parasuraman and Riley (1997), users of automation may feel they are building trust with designers other than automations during interactions, so it is also probable that humans perceive morality-related violations differently and make different violation attributions during HRI. Sebo et al. (2019) manipulated their robot to explicitly articulate the reasons (e.g., “Oh no! I hit the wrong button” and “Yes! You’re immobilized”) of trust violations to ensure that interactants perceived them as certain kind of violations. Since, however, machines do not always frame the intensions behind violations so clearly in real life, it remains unclear whether interactants in HMC perceive competence- or integrity-based violations in the same manner as they do in interpersonal communication. Therefore, applying the rules from human-to-human trust repair straight to robot-to-human trust repair may elude some crucial insights concerning these interactions.

Robotic Failures

“Failures” and “errors” are often used interchangeably in HMC research together with “faults”. In the present study, they are approached as overlapping but distinctive terms. First, the term “failures” refers to “a degraded state of ability which causes the behavior or service being performed by the system to deviate from the ideal, normal, or correct functionality” (Brooks, 2017, p.9), emphasizing violations of interactants’ subjective expectations. Second, the word “errors” is a more technical term, encompassing “system states (electrical, logical, or mechanical) that can lead to a failure” (Honig & Oron-Gilad, 2018). Third, the term, “faults,” is defined as lower-order sources of errors (Honig & Oron-Gilad, 2018). Errors might cause failures, if noticed and perceived by human users as failures, but failures do not necessarily result from errors—misperceptions and incompatible designer principles can also engender failures. Failures can be both competence- and integrity-based, generated unintentionally because of system errors or intentionally because of gaps between designer and user goals. Starting from the division made by Kim et al. (2004), failures caused by system errors are apparently competence-based from the robotic end, and they are caused by unintended system inability preventing robots from producing correct output and executing human commands accurately.

Taxonomies of Failures

Previously, numerous failure typologies were constructed to explain errors emerging in HMC introduced by humans, robots, and the environment. Based on the locus of faults, they were categorized into (1) physical, human-made, design and interaction faults (Laprie, 1985), (2) information acquisition, information analysis, decision/action selection, and action implementation (Parasuraman et al., 2000), (3) interaction, algorithms/methods, software design/implementation, and hardware failures (Carlson & Murphy, 2005), (4) interactions, algorithms, software, and hardware faults (Steinbauer, 2012) and (5) communication failures and processing failures (Brooks, 2017). Based on situations of expectation violations, Giuliani et al. (2015) distinguished failures by technical failures and social norm violations. Based on mechanisms of errors, Skitka et al. (2000) emphasized omission and commission errors. Based on combinations of failures, Ferrell (1994) organized robotic failures into individual, concurrent, and accumulative failures. Based on severity of aftereffect, they were classified into (1) benign failures and catastrophic failures (Laprie, 1995), (2) non-critical, repairable/compensable, and terminal failures (Carlson & Murphy, 2005). Based on recoverability of failures, they could be divided into anticipated, exceptional, and unrecoverable errors (Ross et al., 2004); and based on cross-contextual applicability, they were identified as high, medium, and low relevancy failures (Honig & Oron-Gilad, 2018).

Despite the appreciable amount of efforts devoted to taxonomy constructions, these categorizations were rarely integrated into error-focused experimental designs, especially for applied research. For example, Kohn et al. (2018) experimented with six common failures of self-driving cars (e.g., crashes, wrong U-turns, delayed starts) on the basis of existing empirical findings. The advantage of employing individual errors originating from usage and practice lies in the instant applicability of such results to relevant systems, but the disadvantages are also conspicuous: the underlying mechanisms causing such differences remain obscure given the limited depth of data interpretation, and the findings cannot be easily applied to other contexts because of the medium or low relevancy.

Though scant, accessible literature that compares theoretically justified error or failure types has shown that they most likely have distinct influences over trust. The error types attracting the greatest amount of scholarly attention so far are commission and omission errors, and this line of research since the early 2000s has mainly scrutinized the effects of false-alarms and misses in automation systems (e.g., Chancey et al., 2015; Davenport & Bustamante, 2010; Dixon, 2007; Dixon & Wickens, 2003; 2004; 2006; Geels-Blairet al., 2013; Johnson et al., 2004; Levinthal & Wickens, 2006; Madhavan et al., 2006; Rice, 2009; Rovira & Parasuraman, 2010; Sanchez, 2006; Sanchez et al., 2004). These experiments led to mixed findings: some indicated misses had more negative valence (e.g., Davenport & Bustamante, 2010; Dixon & Wickens, 2003; Sanchez, 2006), but others suggested false alarms were worse (e.g., Johnson et al., 2004), with some viewing both as equally destructive (e.g., Madhavan et al., 2006; Rovira & Parasuraman, 2010). Primarily, researchers differentiated two the types of errors based on their relationships with two dimensions of trust behavior, compliance and reliance, with minor references to workload, salience of errors, and outcome values (Sanchez, 2006).

To merge the gap in literature, Hoff and Bashir (2015) commented that the contradictions in previous finding might have been caused by different consequences of errors across systems—a false alarm of a carbon monoxide detector might simply be an annoyance, yet a miss could lead to casualties—meaning that future research looking into error types must cautiously control predicted outcomes caused by different kinds of errors. Apart from errors of commission and omission, Flook et al. (2019) investigated technical and decision-level failures in HRI, and their findings showed that two failure types had similar effects, refuting the hypothesis that socio-level failures dampen participants' trust more seriously because technical failures are considered to be easier to amend while recognition of social signals is perceived to be a higher-rank capability. Overall, the connection between existing failure typologies and empirical research has been tenuous, and future research needs to explore failure effects with more refined theoretical considerations.

Trust Repair Based on Failure Types

Trust repair research related to different violation types is relatively a new topic in HMC, and there is a noteworthy HAI framework formulated by Marinaccio et al. (2015). This framework connects the aforementioned discoveries from organizational trust literature (Kim et al., 2004; 2013) and human error typology from Reason (1990), surmising that the same interactions between two violation types and two repair methods also manifest themselves in human-automation relationships. In Reason's classification (1990), error was utilized as a generic term (i.e., "all occasions in which a sequence of planned mental or physical activities fail to achieve its intended outcomes," p.9), synonymous with "failures" in the present study. Thus, "violations" (i.e., intentional commission of an error), a type of failures that results from intended errors as a form of integrity-based violations (Marinaccio et al., 2015), do not count as "errors" in the present study based on the given definition. "Mistakes", on the other hand, allude to decision-level failures aggregating prior errors and appropriateness of entire system designs, which are not elevated to the same level as the other two sorts of technical errors, "slips" and "lapses", so it is determined that leaving out mistakes in the experimental design of this study would be reasonable. Table 1 below includes the relationships Marinaccio et al. (2015) drew between slips, lapses, mistakes, and violations defined by Reason (1990) and trust repair typology delineated by Kim et al. (2013). Quinn (2018) and Sebo et al. (2019) have observed partial and full support to the interactions between violation types and effective repair, suggesting the findings about trust repair in human-to-human communication remain instructive in HMC.

Table 1

Trust Repair Framework Proposed by Marinaccio et al. (2015)

Error Type (Reason, 1990)	Examples	Violation Types (Kim et al., 2013)	Effective Repair (Kim et al., 2013)
Slips – Errors of commission – when an intended action is wrongly executed	Flipping the wrong switch on an IV pump	Integrity-based	Denial
Lapses – Errors of omission – resulting in failure to carry out the action	Forgetting to administer medication	Competence-based if due to memory failure, integrity-based if due to attention failure	Context-dependent
Mistakes – Errors of planning or judgment	Prescribing an incorrect dosage	Competence-based	Apology
Violations – Intentional commission of an error	Prescribing an inappropriate medication because of sponsor loyalty	Integrity-based	Denial

Three categories of technical failures. Reason's (1990) identification of errors largely relies on recognition of the stage in which errors occur, and this process-centered view might not fully reflect interactants' subjective perceptions of failures. Departing from three origins of failures, planning, storage, and execution, Reason (1990) deemed slips and lapses to be execution-based and/or storage-based deficiencies and mistakes to be planning-based deficiencies. But untrained users do not necessarily probe into the mechanisms underlying error occurrence since symptoms and sources of system failures are often hard to comprehend even for experts in the field (Honig & Oron-Gilad, 2018); instead, they make judgments about qualities of wrongness predominantly based on available output. Are responses successfully delivered? Are they correct? If they are incorrect, in which way are they wrong? What or who do they think should take the blame then? Lacking professional knowledge, interactants probably do not consider whether failures are caused by memory lapses or attention failures when coming across errors of omissions, for example. As discussed above, human errors and robotic errors are probably perceived differently because of ontological differences, so the demarcation between competence- and integrity-based violations based on interpersonal principles might not be exactly the same for HRI—flipping the wrong switch and delivering an incorrect dosage might be both considered incompetent, though Marinaccio et al. (2015) attributed them to different types of trust violations based on different causes.

The same concern pertains to other extant failure taxonomies surrounding locus of faults that from the eye of human users, precise origins of errors caused by their robotic counterparts are little known. To resolve such conflicts and overcome shortcomings, basic error types (i.e., logic, sematic, syntax errors) in computer science (McCall & Kölling, 2014) can be adopted for developing an execution-centered failure categorization and investigating technical failures in some more details (Table 2). Take, for instance, a hypothetical task in which a robot is instructed to build a toy tower. Logic errors are termed as errors causing machines to produce relevant but incorrect output, which covers a part of slips, such as retrieving four building blocks when asked to bring three; sematic errors refer to errors yielding completely irrelevant output inappropriate in the given context, which blankets all non-logic slips, such as singing a song when required to pick up a stick; syntax errors are essentially errors of omission, in which cases machines fail to run programs, such as giving no responses to human commands. The three categories of errors will lead to three types of failures when interactants heed flawed robotic output, so they are labeled according to their error origins as logic failures, semantic failures, and syntax failures.

Such failures are common in HRI. For instance, two types of failures naturally occurred in Pino et al.'s (2020) study when a NAO robot served as a trainer for cognitively impaired elderly—their NAO sometimes incorrectly evaluated participants responses and demonstrated logic failures (e.g., judging "10" to be the right answer for "what is answer to 5+8?") and syntax failures (e.g., not responding to instructions); command recognition errors also frequently happen to robotics (Iio et al., 2020), which might lead to semantic failures (e.g., misunderstanding commands and exercising irrelevant action). From vending machines and printers to personal digital assistants and chatbots, this outcome-centered typology transcends the technic divisions of hardware and software and is applicable to most existing machines, including automations and virtual systems.

Table 2

Trust Repair Framework for Technical Errors Only

Failure Types	Examples	Violation Types (Kim et al., 2013)	Effective Repair (Kim et al., 2013)
Logic failures – resulting in relevant but wrong action	Flipping the wrong switch on an IV pump	Competence-based	Apology
Semantic failures – resulting in irrelevant or meaningless action	Reduce dosage on record when asked to print out a prescription	Competence-based	Apology
Syntax failures – resulting in failures to carry out the action	Forgetting to administer medication	Competence-based	Apology

The approach to violation types in this study is fundamentally different from Sebo et al.'s (2019) in that failures are not famed as competence- or integrity-based. Since humans less frequently make moral-related attributions to machines, it is deducted that these failures would all be subjectively conceptualized as competence-based violations instead of integrity-based violations. This speculation is different from Marinaccio et al.'s (2015) propositions that humans would perceive the robotic motivations behind slips, including both logic and semantic failures, to be integrity-based, as they do in interpersonal communication. Quinn (2018) and Sebo et al. (2019) have verified in HMC that apology with internal attribution rather than denial repairs trust more effectively for competence-based violations, and vice versa for integrity-based violations. If participants take all three failure types as competence-based violations, their trust would be better recovered with apology with internal attribution than denial.

> **H1.** After failure occurrence, participants' trust in the robot will be repaired more successfully when it repairs trust with internal-attribution apology rather than denial for (a) logic, (b) semantic, and (c) syntax failures.

Semantic failures may appear more objectionable to humans than logic failures because in semantic errors robots fail to interpret input at the very beginning, while robots appear to understand interactants' input to some degree in logic errors. So even though both failures are competence-based, outcomes of semantic errors may be viewed more severely and negatively affect evaluations of the robots. But for logic errors, they may appear harder to detect and therefore elicit more negativity than semantic failures. Meanwhile, there also exists the possibility that humans are prone to believe failures are caused by human errors when robots give completely meaningless output, which makes logic failures more negative than semantic failures. Nevertheless, more empirical support to this deduction is needed. Considering prior studies also presented mixed findings regarding miss- and false-prone errors, which are essentially semantic and logic failures, it is hard to predict the magnitude of violations of different failure category. To explore the nature of these violations, the following research questions were scrutinized:

RQ1. How will different types of failures affect (a) perceived competence, (b) perceived integrity, (c) competence-based post-interaction trust, (d) integrity-based post-interaction trust, and (e) perceived severity of violations, when no trust repair is implemented?

RQ2. Which type of failures will exert the strongest negative effects over participants' post-interaction trust in robot, regardless of repair methods?

Blame Attributions in Trust Repair

Trust Repair as Attribution Manipulation

Attribution theory is one of the most salient theoretical perspectives in trust repair research (Lewicki & Brinsfield, 2017; Tomlinson & Myer, 2009), which has been applied in multiple studies to explain the effects of trust repair methods (e.g., Bansal et al., 2015; Goles et al., 2009; Quinn, 2018). As a building block of contemporary psychology, attribution theory has greatly advanced our understanding since late 1950s of how people attribute causes of events and respond accordingly (Weiner, 2008). The fundamental distinction Heider (1958) propounded over people's assigned explanations of behavior and events is locus (i.e., whether perceived causes are located in external situational factors or the actor's internal qualities). Later, the theory was further elaborated with two additional dimensions: stability (i.e., whether perceived causes are fluctuant or constant) and controllability (i.e., whether perceived control of reinforcement is external or internal), and these dimensions are closely linked to individuals' expectancy changes and emotional responses (Weiner, 1985).

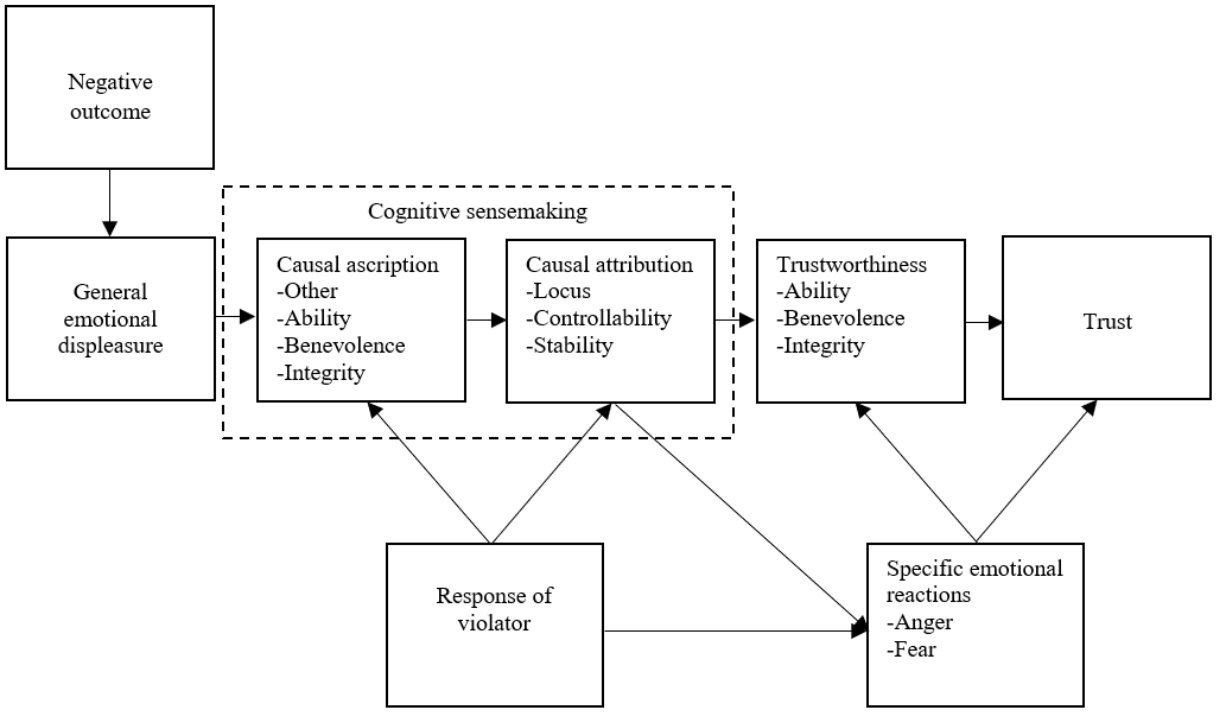
Attributions play a pivotal role in trust repair (Dirks et al., 2009). Benevolent attributions for failures, the ones that are more external, unstable, and uncontrollable, can stimulate more favorable outcomes and encourage forgiveness, while internal, stable, and controllable attributions lead to more negativity in failure assessment (e.g., Korsgaard et al., 2002; Shaw et al., 2003; Stouten et al., 2006; Takaku, 2001). According to Weiner's (1985) typology, individuals' poor aptitudes might be perceived as caused by internal, stable, and uncontrollable reasons, while immorality might be assigned with internal, stable, and controllable attributions, which offers an explanation of why integrity-based violations are taken more seriously as trust violations than competence-based violations. Based on the general findings about blame attributions, the current study proposed the following hypothesis:

H2. After failure occurrence, higher levels of trust will be assigned to robots when more (a) external, (b) unstable, and (c) uncontrollable causal attributions are made, regardless of failure types and repair methods.

Different trust repair strategies can be approached as different ways of manipulating causal attributions. Based on the past research in trust repair, Weiner's (1985) attribution literature, and Mayer et al.'s (1995) model of organizational trust, Tomlinson and Mryer (2009) proposed a model for trust repair concerning attribution manipulation (see Figure 1). However, because the present study only examines integrity- and competence-based violations, the component of benevolence is excluded from the discussion. The latent logic in trust repair is that one end of each attribution continuum tilts the other—for instance, if one makes more external attributions in the case, he or she will naturally reduce internal attributions—so that trustees can make more external attributions in trust repair attempts to decrease trustors' internal attributions of guilt (Crant & Bateman, 1993). This hence alleviates the negative effects of violations on perceived trustworthiness by ruling out the notion that that failures are caused by certain deficiencies in ability or integrity (Tomlinson & Mryer, 2009).

Figure 1

Attribution Model of Trust Repair from Tomlinson & Myer (2009)



Finally, denial and apology can both contribute to effective trust repair by predisposing trustors to make more benevolent attributions from the lens of attribution theory. Based on previous findings, Tomlinson and Mryer (2009) proposed that damaged perceptions of competence can be repaired with attributions to external factors as well as unstable and/or uncontrollable forms of abilities, while integrity-based violations can benefit from external or unstable internal causes. Denial (e.g., "It was not my fault") asserts external attributions (Baker et al., 2018; Quinn, 2018) while apology, defined by Kim et al. (2004), typically weakens stability attributions by portraying unstable statuses of aptitude (e.g., "I promise to do better in the future"), although provoking internal attributions by accepting the responsibilities (e.g., "Sorry, it was my fault").

H3. After failure occurrence, more (a) unstable and (b) uncontrollable attributions will be made when the robot repairs trust with either internal- or external-attribution apology than when it takes no action, regardless of failure types.

H4. After failure occurrence, more external attributions will be made when the robot repairs trust with denial than when it takes no action, regardless of failure types.

Apology with Internal and External Attributions

Beyond the basic division between apology and denial, Tomlinson et al. (2004) and Kim et al. (2006) discussed additional variations of apology. After accepting the responsibility for violations, trustees have two possible ways of framing locus of causes: they can either make external attributions (e.g., "Sorry, the question was phrased too ambiguously") or internal attributions (e.g., "Sorry, I was too timid to ask questions") to explain their failures. While an array of research has marked positive effects of external attributions, some studies highlighted their potential risks. For example, finding excuses can be perceived to be deceptive, self-absorbed, and ineffectual (Schlenker et al., 2001), thus diminishing trustors' willingness to reconcile (Tomlinson et al., 2004). Since people do not like lying robots (Wijnen et al., 2017), external attributions can be counterproductive when trustors are not convinced of robots' innocence. Given both studies found apology with internal attributions rehabilitate trust better for competence-based violations, it is also expected to acquire similar findings in the HRI settings:

H5. After failure occurrence, participants' trust in the robot will be repaired more successfully when robots repair trust using apology with internal attribution than with external attributions, regardless of failure types.

Kim et al. (2006) did not compare apology with external attributions with denial. Following Kim et al.'s (2004) argument that positive information outweighs negative information in competence-based violations, apology with both external and internal attributions should provide more positivity than denial. Even if an apology with external attributions accepts only partial responsibility for the failures, the conveyed remorse and sincerity in apology with external attributions should still outperform denial as a trust repair method for competence-based violations.

H6. After failure occurrence, participants' trust in the robot will be repaired more successfully when the robot repairs trust using apology with external attribution than using denial, regardless of failure types.

No previous study has ever compared denial with no repair. Based on previous research on human-to- human communication, denial is likely to be perceived as repulsive and deceptive under competence-based failures, so the present study hypothesizes it will be more harmful than taking no action.

H7. After failure occurrence, participants' trust in the robot will be higher when the robot takes no action than repairing trust with denial, regardless of failure types.

Eventually, the research question concerning possible the interaction effects between failure types and repair methods is generated:

RQ3. After failure occurrence, how will failure types and trust repair methods interact with each other concerning participants' trust in the robot?

Implicit Theories of Moral Responsibility

Apart from the classic attribution theory, there is another important framework that illustrates the effects of people's beliefs about human attributes on their judgements and reactions in blame attribution (Chiu et al., 1997; Dweck et al., 1995; Gervey et al., 1999), implicit theories of moral responsibility. The theories posit people tend to explain actions with fixed traits if they believe personal traits are nonmalleable (entity theory), while they are inclined decipher causes in terms of situational factors if they hold the opinion that human attributes are dynamic malleable (incremental theory) (Chiu et al., 1997; Dweck et al., 1995; Gervey et al., 1999). From the viewpoint of implicit theories, Kam (2009) proposed that implicit theory beliefs actively shape effectiveness of trust repair outcomes. According to Kam (2019), individuals with entity mindsets are more likely to make internal, stable, and controllable attributions, compared with individuals with incremental beliefs; trust violations should have more negative impacts on individuals with entity orientation in contrast to individuals with incremental orientation, whereas trust repairs will be less successful for entity-oriented individuals, with slower trust recovery. Therefore, the present study also incorporated participants' entity beliefs as an essential covariate in HRI trust repairs.

Method

Participants

The current study's sample consisted of 330 undergraduate students enrolled in communication courses at a major Southwestern U.S. university, with 39 incomplete responses excluded from final analysis. Data collection lasted from February 2nd, 2021 to May 6th, 2021, approved by the Institutional Review Board (IRB) of the university. Their age ranged from 18 to 27 (M = 20.07, SD = 1.48), with 190 being female (65.29%), 99 being male (34.02%), and 2 being other (0.7%). Nationality-wise, most of them were Americans (n = 245, 84.19%). For ethnicity, Caucasian/white dominated the sample (n = 210, 72.16%), followed by Latino/Hispanic (n = 22, 7.56%), mixed ethnicity (n = 16, 5.50%), Asian (n = 15, 5.15%), Black or African American (n = 14, 4.81%), Native Indian or Alaska native (n = 6, 2.06%), other (n = 4, 1.37%), and Native Hawaiian or Pacific islander (n = 1, 0.34%).

Procedure

The robot used in the present study was a NAO robot developed by SoftBank Robotics (2020), which commonly serves educational, research, business, and healthcare purposes. It had been previously programmed to perform tasks taking various roles, such as personal/service assistants (e.g., Vega et al., 2019), human/robot team members (e.g., Sebo et al., 2018), healthcare/therapy assistants (e.g., Shamsuddin et al., 2012), social robots (e.g., Pelikan & Broth, 2016), and instructors (e.g., Park et al., 2011). In the present study, NAO was portrayed as a healthcare assistant capable of providing information about patients' prescriptions.

The study took a between-within subject design under thirteen conditions (3 failure types × 4 repair attempts + 1 control). Participants first provided their demographic information (i.e., age, sex, nationality, ethnicity) and answered questions asking their propensity to trust robotics (i.e., the general tendency to trust robots) and entity beliefs (i.e., to which extent individuals believe personal traits are fixed). Based on their sex, they were thereafter directed to a set of pre-recorded videos with an either male or female interactant voice—males were matched with the male interactant voice while females were matched with the female interactant voice; when identified as other gender, participants were randomly assigned to either one of the conditions. Before viewing these videos recorded from a first-person perspective, participants were presented with some basic information about real-world applications of NAO robots and were asked to imagine that they were actually living through these episodes in provided videos (adapted from Smith & Lazarus, 1993): "Imagine that you were going through this interaction yourself as the person interacting with NAO and answer the following questions".

After watching a brief introduction video from a NAO robot that presented itself as a healthcare assistant and engaged in social conversations with the interactant, participants were randomly assigned to three of the thirteen conditions with randomized orders, so eventually there were 743 observations, excluding the ones failing to pass the attention verifications. Even though such observations were not completely independent from one another as each participant was assigned to three conditions, the randomization of condition combinations and presentation orders mitigated this violation. Hence, the observations were just partial violations of independent observations, which was acceptable because general linear models are robust enough against such issues.

In the control condition, participants were instructed to report their trust, perceived competence, and integrity of the robot after viewing an interaction episode in which the NAO robot answered the interactant's questions concerning a given prescription perfectly. In each experimental condition video, NAO demonstrated one of the three types of performance failures (i.e., logic, semantic, and syntax) and either made no repair attempts or repaired trust with one of the three strategies (i.e., internal-attribution apology, external-attribution apology, and denial). And after they finished watching each video, an attention check was implemented to examine whether participants noticed those performance failures with one closed-ended question (i.e., "Recall the interaction you just saw. Did the NAO robot make any mistake(s)?") and one open-ended question (i.e., "Recall the interaction you just saw. Please briefly describe what kind of mistake(s) the NAO robot made, if any; if the robot did not make any mistake(s), please answer with 'NA.'"). Repeated measures followed every video to capture participants' perceived competence and integrity, trust in the robot, severity of trust violations, and causal attributions. Another attention check (i.e., "This is an attention verification question. Please answer with…to the question") was embedded amongst the measures for each condition.

Measures

Propensity to Trust

Conceptualized as individuals' stable traits, general disposition to trust robots is associated with usage beliefs and intents (Merritt & Ilgen, 2008). Participants' propensity to trust was measured with six 5-point Likert-type scale items (1 = *Strongly agree*, 5 = *Strongly disagree*) adapted from the Propensity to Trust Technology Scale developed by Jessup et al. (2019) for the present research context. The sample items included "generally, I trust robots", "I think it's a good idea to rely on robots for help", and "I don't trust the information I get from robots" (Reverse coded). The scale has been previously adapted for human-to-automation trust and reached good internal reliability, Cronbach's $\alpha = .84$ (Jessup et al., 2019). In the present study, the scale reliability was acceptable, Cronbach's $\alpha = .71$, and the average score of the scale was utilized for further analysis ($M = 3.26$, $SD = 0.76$).

Entity Beliefs

Participants' entity beliefs (i.e., to which extent individuals believe personal traits are nonmalleable) were measured with six 7-point Likert-type scale items (1 = *Strongly disagree*, 7 = *Strongly agree*) developed by Dweck et al. (1995), which originally included nine items measuring three dimensions (i.e., intelligence, morality, and world). Because the present study focused on competence- and integrity-based violations, only the first two dimensions (e.g., intelligence and morality) of the measures were included. Three items assessed participants' entity beliefs on human intelligence: "A person has a certain amount of intelligence and he/she really can't do much to change it", "A person's intelligence is something about him/her that he/she can't change very much", and "A person can learn new things, but he/she can't really change his/her basic intelligence"/ Three items were used to measure entity mindsets on morality: “A person's moral character is something very basic about them, and it can't be changed much”, “Whether a person is responsible or sincere or not is deeply ingrained in their personality. It cannot be changed very much”, and “There is not much that can be done to change a person's moral traits (e.g., conscientiousness, uprightness and honesty)”. The subscales measuring intelligence-based (Cronbach’s α = .88) and morality-based entity beliefs (Cronbach’s α = .78) acquired good internal reliability. The confirmatory analysis showed the global goodness of fit indices from the initial bifactor model (RMSEA = .10, CFI = .972, SRMR = .04) met Hu and Bentler’s criteria (1999) except for RMSEA: RMSEA ≤ .06, CFI ≥ .95, SRMR ≤ .08. After allowing significant error covariances between the items belonging to the same dimension, the model fit was improved (RMSEA = .05, CFI = .997, SRMR = .01). Two dimensions were highly correlated ($r = .72$, $p < .001$), and the average score of the entire scale was utilized for further analysis ($M = 2.93$, $SD = 1.21$).

Perceived Competence and Integrity

Four items were adapted from the six 7-point items (1= *Strongly disagree*, 7 = *Strongly agree*; Cronbach’s α = .87; $M = 4.60$, $SD = 1.23$) measuring perceived ability developed by Mayer and Davis (1999), three of which were later tailored by Kim et al. (2004) to capture perceived competence in the robot. Since the original items were designed to measure organizational trust, two items inapplicable under the current context were dropped and the wordings were modified to fit the purpose of this study. The sample statements included “The robot is very capable of performing its job” and “The robot is successful at things it tries to do”. In a similar manner, another four 7-point items (1= *Strongly disagree*, 7 = *Strongly agree*; Cronbach’s α = .82; $M = 4.71$, $SD = 1.17$) measuring perceived integrity were adapted from the scale that Mayer and Davis (1999) and Kim et al. (2004) used to measure perceptions of integrity, including “The robot sticks to its word” and “Sound principles seem to guide the robot’s behavior”. Perceived competence was highly correlated with perceived integrity ($r = .64$, $p < .001$).

Post-Interaction HRI Trust

Seven 11-point items (1 = *0%*, 11 = *100%*) extracted from the HRI-Trust Perception Scale (Schaefer, 2013) with the highest Content Validity Rations (CVR) values[1] from the original study evaluated competence-based trust, and the average score was used for further analysis, $M = 7.73$, $SD = 1.93$. Seven 7-point Likert items (1 = *Not true at all*, 7 = *Very much true*) adapted from Jian et al.’s (2000) Checklist for Trust Between People and Automation Scale were used to assess participants’ post-interaction trust based on perceived integrity of the NAO robot, and the mean score was retained for further analysis, $M = 4.62$, $SD = 1.05$. The reason why these two scales were simultaneously employed was because HRI-based trust is a multidimensional construct, and the former emphasized robots’ competence relatively more (e.g., “What % of the time will this robot function successfully?”), whereas the latter shed more light on integrity- and benevolence-based trust (e.g., “The robot has integrity.”). While the latter had proven to be reliable as a classic measurement instrument in the field, the former was a relatively new scale. The first scale reached high internal reliability (Cronbach’s $\alpha = .96$), and the reliability of second one was also acceptable (Cronbach’s $\alpha = .76$). Competence-based post-interaction trust was positively correlated with integrity-based post-interaction trust ($r = .54$, $p < .001$). Besides, competence-based post-interaction trust was positively associated with both perceived competence ($r = .72$, $p < .001$) and integrity ($r = .55$, $p < .001$), and integrity-based

[1] $CVR = (n_e - N/2) / (N/2)$

post-interaction trust was also highly correlated with perceived competence (r = .58, p < .001) and perceived integrity (r = .59, p < .001).

Severity of Failures

Three 5-point items from Weun et al. (2004) assessed perceived severity of technical failures. The scale was originally constructed to evaluate service failure severity, and it achieved composite reliability of .93 in the initial study. The items were "If this problem were really happening to me, I would consider the problem to be… (1= *Not very severe*, 5 = *Very severe*)", "If this problem were really happening to me, it would make me feel… (1 = *Not very angry*, 5 = *Very angry*)", and "If this problem were really happening to me, it would be unpleasant to me (1= *Strongly disagree*, 5 = *Strongly agree*)." The reliability of this scale was relatively low in the present study, Cronbach's α = .55, so the first item was deleted to promote scale reliability, Cronbach's α = .70. The average score of two remaining items was calculated for further analysis, M = 2.92, SD = 0.90. Perceived severity of failures was found to be negatively associated with perceived competence (r = -.48, p < .001), perceived integrity (r = -.35, p < .001), post-interaction competence-based trust (r = -.40, p < .001), and integrity-based trust (r = -.51, p < .001).

Causal Attributions

Causal attributions of robotic failures were measured with twelve 9-point bipolar items from the Revised Causal Dimension Scale (CDSII; McAuley et al., 1992), and each subscale contained three items, with wording of the items slightly adjusted to match the context of this study. This scale was designed by McAuley et al. (1992) in a way that attributions of controllability was reflected by two discriminant subdimensions: external (i.e., whether the cause can be controlled by the NAO robot in the videos) and personal control (i.e., whether the cause of robotic failures is under the control of the human interactant). As a result, the scale contained four subscales altogether (i.e., locus of causality, external control, stability, and personal control; see Table 3 for scale reliability, descriptive statistics, and bivariate correlations). The higher scores in each subscale stood for the higher levels of perceived external locus of causality, non-external control, instability, and non-personal control. The confirmatory analysis indicated the initial model fit did not meet the given criteria (RMSEA ≤ .06, CFI ≥ .95, SRMR ≤ .08.) proposed by Hu and Bentler's (1999): RMSEA = .08, CFI = .92, SRMR = .08. According to DeVellis (2016), a strong path coefficient should be .65 and above, so one indicator from the locus subscale and the other from the stability subscale were dropped, which significantly improved the model: RMSEA = .06, CFI = .97, SRMR = .05.

Table 3

Scale Reliability, Descriptive Statistics and Correlations for Subdimensions of CDSII

Variable	Cronbach's α	*M*	*SD*	1	2	3	4
1. Locus	.71	6.66	1.72	-			
2. External control	.76	6.35	2.70	$-.16^{*}$	-		
3. Stability	.65	6.37	1.72	$.24^{*}$	$-.33^{*}$	-	
4. Personal control	.88	5.62	2.11	$.43^{*}$	$.09^{*}$	$-.10^{*}$	-

* $p < .01$

Results

Mean substitutions were implemented as the remedy for missing data. Initially, all variables were normally distributed based on the criteria suggested by Osborne (2003) that skewness and kurtosis with absolute values smaller than 1 should not raise concern. To test effectiveness of the manipulation, A multivariate analysis of covariance (MANCOVA) was first conducted to examine whether the experimental conditions significantly differed from the failure-free condition after removing five multivariate outliers, $p < .001$, with perceived competence and perceived integrity entered as dependent variables as well as propensity to trust robots and entity beliefs entered as covariates. Because propensity to trust was a significant covariate, $p < .01$, whereas entity beliefs was not, $p = .53$, the MANCOVA model was reconstructed after excluding entity beliefs.

The assumption of homogeneity of covariance was met based on the cutoff value, .01, suggested by Tabachnick and Fidell (2007), Box's $M = 57.09$, $F(36, 980133) = 1.56$, $p = .02$. Levene's test showed error variances were equal for perceived integrity, $F(12, 725) = 0.69$, $p = .76$, but not for perceived competence, $F(12, 725) = 3.01$, $p = .0004$, at the .01 level—this more stringent cutoff value proposed by Tabachnick and Fidell (2007) was accepted by the present study because model robustness is expected. The significant differences across groups were identified at the omnibus level, Wilks' Lambda (λ) = .84, $F(24, 1446) = 5.58$, $p < .001$, partial $\eta^2 = .09$, with propensity to trust being a significant covariate, Wilks' Lambda (λ) = .98, $F(2, 723) = 6.02$, $p < .001$, partial $\eta^2 = .02$. Pairwise comparisons indicated the control group had significantly higher levels of perceived competence than every experimental group at the .001 level, and also higher levels of perceived integrity than all experimental groups at the .01 level, except for logic failures with internal-attribution apology, semantic failures with internal-attribution apology and no repair, and syntax failures with internal/external apology and no repair. The results showed the experimental groups generally elicited lower levels of perceived competence and sometimes lower levels of perceived integrity than the failure-free control group, indicating the manipulation was effective on the baseline level.

Results of H1, RQ1, and RQ2

H1 predicted that apology with internal attribution would be more effective as a trust repair method than denial, because (a) logic, (b) semantic, and (c) syntax failures are all competence-based trust violations. For H1a, a MANCOVA test was conducted to first test the effects of internal-attribution apology vs. denial on both types of post-interaction trust under the category of logic failures, with propensity to trust robots and entity beliefs as covariates. Since both propensity to trust, $p = .48$, and entity beliefs, $p = .69$, were insignificant covariates, the model was revised into a multivariate analysis of variance (MANOVA) model.

Error variances were equal across groups at the .01 level for integrity-based trust, $F(1, 123) = 2.81$, $p = .10$, but not both competence-based trust, $F(1, 123) = 8.56$, $p = .004$, according to Levene's tests based on means, which might bias the test results; the assumption of equality of covariance was met, Box's $M = 6.37$, $F(3, 2942853) = 2.09$, $p = .10$. The group differences were significant on the multivariate level, Wilks' Lambda (λ) = .92, $F(2, 122) = 5.29$, $p < .01$, partial $\eta^2 = .08$. The mean difference in competence-based trust, $F(1, 123) = 6.05$, $p < .05$, partial $\eta^2 = .05$, was significant, and so was the mean difference in integrity-based trust, $F(1, 123) = 9.43$, $p < .01$, partial $\eta^2 = .07$ (see the first two rows in Table 4 and 5 for mean differences). Therefore, H1a was supported.

A similar MANCOVA test was performed under the condition of semantic failures to examine H1b, which was revised into a MANOVA model after excluding two insignificant covariates, propensity to trust, $p = .08$, and entity beliefs, $p = .41$. The assumption of equality of covariances was met on the .01 level, Box's M = 8.21, $F(3, 2433191) = 2.68$, $p = .05$, and error variances were equal across groups at the .01 level based on Levene's tests for both competence-based trust, $F(1,114) = 6.03$, $p = .02$, and integrity-based trust, $F(1,114) = 0.21$, $p = .65$. The group effects were significant at the omnibus level, Wilks' Lambda (λ) = .94, $F(2,113) = 3.83$, $p < .05$, partial $\eta^2 = .06$. However, univariate tests showed that the mean difference for competence-based trust was insignificant, $F(1, 114) = 2.16$, $p = .15$, partial $\eta^2 = .02$, but significant for integrity-based trust, $F(1, 114) = 7.73$, $p < .01$, partial $\eta^2 = .06$ (see the third and fourth rows in Table 4 and 5 for mean estimates and differences). Since competence and integrity are two aspects of post-interaction trust, H1b was partially supported.

Table 4

Means and Mean Estimates for Internal-Attribution Apology vs. Denial Under Three Failure Types

Dependent Variable	Failure	Repair	Mean	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Competence-Based Trust	Logic	Apology (Internal)	8.09	0.23	7.63	8.56
		Denial	7.29	0.23	6.83	7.74
Integrity-Based Trust	Logic	Apology (Internal)	4.66	0.14	4.38	4.94
		Denial	4.06	0.14	3.79	4.33
Competence-Based Trust	Semantic	Apology (Internal)	7.54	0.25	7.05	8.03
		Denial	7.03	0.24	6.55	7.51
Integrity-Based Trust	Semantic	Apology (Internal)	4.76	0.12	4.52	4.99
		Denial	4.29	0.12	4.06	4.52
Competence-Based Trust	Syntax	Apology (Internal)	7.56	0.35	6.85	8.26
		Denial	7.04	0.35	6.33	7.74
Integrity-Based Trust	Syntax	Apology (Internal)	4.72	0.18	4.36	5.08
		Denial	4.33	0.18	3.97	4.69

Notes. For syntax failures, covariates appearing in the model are evaluated at the following values: propensity to trust = 3.25, entity beliefs = 2.96.

Table 5

Pairwise Comparisons for Internal-Attribution Apology vs. Denial Under Three Failure Types

Dependent Variable	Failure	(I) Repair	(J) Repair	Mean Difference （I-J）	Std. Error	Sig.[b]	95% Confidence Interval for Differences Lower Bound	95% Confidence Interval for Differences Upper Bound
Competence-Based Trust	Logic	Apology (Internal)	Denial	0.81*	0.33	.02	0.16	1.46
Integrity-Based Trust	Logic	Apology (Internal)	Denial	0.60**	0.20	.003	0.21	0.99
Competence-Based Trust	Semantic	Apology (Internal)	Denial	0.51	0.35	.15	-0.18	1.20
Integrity-Based Trust	Semantic	Apology (Internal)	Denial	0.46**	0.17	.006	0.13	0.79
Competence-Based Trust	Syntax	Apology (Internal)	Denial	-0.02	0.18	.96	-0.64	0.61
Integrity-Based Trust	Syntax	Apology (Internal)	Denial	0.36	.18	.06	-0.01	0.72

* The mean difference is significant at the .05 level.
** The mean difference is significant at the .01 level.
[b] Adjustment for multiple comparisons: Bonferroni.

Following a similar procedure, another MANCOVA was conducted to investigate syntax failures. Covariances were equal across groups, Box's M = 2.00, $F(3, 2403279) = 0.65$, $p = .58$, and error variances were equal between groups for both competence-based trust, $F(1, 115) = 0.12$, $p = .73$, and integrity-based trust, $F(1, 115) = 0.09$, $p = .77$, based on Levene's tests. The main effects were not significant on the multivariate level, Wilks' Lambda (λ) = .96, $F(2, 112) = 2.67$, $p = .07$, partial $\eta^2 = .05$. Moreover, there were no significant differences (see the last two rows in Table 4 and 5 for mean estimates and differences) found on either competence-based trust, $F(1, 113) = 0.002$, $p = .96$, partial $\eta^2 = .00002$, or integrity-based trust, $F(1, 113) = 3.74$, $p = .06$, partial $\eta^2 = .03$, after controlling for two covariates, so H1c was not supported. Overall, H1 was partially supported by the test results.

The first research question inquired about the effects of different failure types on participants' perceptions and post-interaction trust under the circumstances in which NAO initiated no trust repair attempts. In response to RQ1a, a one-way analysis of covariance (ANCOVA) was computed in which propensity to trust and entity beliefs were entered as covariates, and perceived competence was entered as a dependent variable. Because the effects of propensity to trust, $p = .39$, and entity beliefs, $p = .99$, were insignificant, an analysis of variance (ANOVA) test was implemented instead. The assumption of homogeneity was met, $F(2, 164) = 0.49$, $p = .61$, and the group differences had insignificant impacts on perceived competence, $F(2, 164) = 0.62$, $p = .54$, partial $\eta^2 = .007$. In response to RQ1b, a similar ANOVA was conducted with perceived integrity as the test variable after excluding two insignificant covariates, propensity to trust, $p = .25$, and entity beliefs, $p = .70$, and the assumption of homogeneity was met, $F(2, 164) = 0.63$, $p = .54$. The grouping effects were also insignificant, $F(2, 164) = 0.34$, $p = .71$, partial $\eta^2 = .004$. For RQ1c, the ANCOVA test was reconducted after excluding an insignificant covariate, propensity to trust, p = .79. Levene's test of equality of error variances was insignificant, $F(2, 164)$ = .60. p = .55. Entity beliefs was a significant covariate in the model, $F(1, 163)$ = 4.50, p < .05, partial η^2 = .03, but the effects of failure types were insignificant in the ANCOVA test, $F(2, 163)$ = 2.35, p = .10, partial η^2 = .03, with competence-based trust as the dependent variable.

For RQ1d, another ANOVA test with integrity-based trust as the dependent variable was run after removing two insignificant covariates, propensity to trust, p = .16, and entity beliefs, p = .15. Levene's test of equality of error variances was insignificant, $F(2, 164)$ =.29, p = .75, and the effects of failure types were also insignificant, $F(2, 164)$ = 2.45, p = .09, partial η^2 = .03. For RQ1e, the effects of propensity to trust, p = .26, and entity beliefs, p = .54, were insignificant in the initial ANCOVA model, so an ANOVA test was performed instead. Error variances were equal between groups based on Levene's test, $F(2, 164)$ = 0.10, p = .91, and the analysis indicated there were no significant group effects on perceived severity of violations, $F(2, 164)$ = 0.60, p = 0.55, partial η^2 = .007. To recapitulate, three types of failures did not elicit significantly different levels of perceived competence, perceived integrity, and post-interaction trust without trust repairs, when compared with one another.

The second research question asked which type of failures generated the strongest negative effects on participants' trust in robots overall. To answer RQ2, twelve experimental conditions were collapsed into three categories of failure types (i.e., logic, semantic, and syntax), and two multivariate outliers were removed, p < .001. Both propensity to trust, p < .001, and entity beliefs, p < .01, were significant covariates in the MANCOVA model. Covariances were equal across groups, Box's M = 3.07, $F(6, 11235870)$ = 0.51, p = .80, and the results of Levene's tests were insignificant at the .01 level for competence-based trust, $F(2, 685)$ = .43, p = .65, and integrity-based trust, $F(2, 685) = 1.18$, $p = .31$. Different failure types were associated with significantly different levels of post-interaction trust in the MANCOVA model with both types of trust (i.e., competence and integrity) as dependent variables and propensity to trust and entity beliefs as covariates, Wilks' Lambda (λ) = .96, $F(4, 1364) = 7.67$, $p < .05$, partial $\eta^2 = .02$.

Test of between subjects effects revealed there were no significant main effects of failure types on competence-based post-interaction trust, $F(1, 683) = 1.85$, $p = .16$, partial $\eta^2 = .005$. But for integrity-based trust, the main effects were significant, $F(1, 683) = 7.99$, $p < .001$, partial $\eta^2 = .02$, with propensity to trust, $F(1, 683) = 17.12$, $p < .001$, partial $\eta^2 = .02$, and entity beliefs, $F(1, 683) = 10.75$, $p < .01$, partial $\eta^2 = .02$, being significant covariates in the model. Pairwise comparisons indicated only the mean difference in integrity-based trust between logic and syntax failures (see Table 6 and 7 for mean estimates and differences) was significant, $p < .001$. Therefore, logical failures were comparatively the most detrimental failure type as far as integrity-based trust was concerned.

Table 6

Mean Estimates for Two Types of Trust Under Three Failure Types

Dependent Variable	Failure	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Competence-Based Trust	Logic	7.69	0.12	7.46	7.92
	Semantic	7.37	0.12	7.13	7.61
	Syntax	7.50	0.12	7.27	7.74
Integrity-Based Trust	Logic	4.35	0.06	4.22	4.47
	Semantic	4.54	0.07	4.41	4.67
	Syntax	4.71	0.07	4.58	4.84

Notes. Covariates appearing in the model are evaluated at the following values: propensity to trust = 3.25, entity beliefs = 2.96.

Table 7

Pairwise Comparisons for Two Types of Trust Under Three Failure Types

Dependent Variable	(I) Failure	(J) Failure	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Differences	
						Lower Bound	Upper Bound
Competence-Based Trust	Logic	Semantic	0.32	0.17	.17	-0.08	0.73
	Logic	Syntax	0.19	0.17	.78	-0.22	0.60
	Semantic	Syntax	-0.13	0.17	1.00	-0.55	0.28
Integrity-Based Trust	Logic	Semantic	-0.19	0.09	.11	-0.41	0.03
	Logic	Syntax	-0.36*	0.09	.0002	-0.58	-0.15
	Semantic	Syntax	-0.17	0.09	.19	-0.40	0.05

* The mean difference is significant at the .001 level.
b Adjustment for multiple comparisons: Bonferroni.

Results of H2, H3, and H4

H2, H3, and H4 investigated the relationships between blame attributions and trust repair. H2 proposed that both competence- and integrity-based trust would be positively associated with (a) external locus of causality, (b) unstable, and (c) uncontrollable causal attributions after error occurrence. In CDSII, controllability of causality was conceptualized with two subdimensions, external control and personal control. Non-external control (i.e., the perception that the failure cause was not under the control of the NAO robot in the videos) should be positively associated with benevolent attributions from the perspective of participants, whereas non-personal control (i.e., the perception that the cause of robotic failures was not under the control of the human actor) should be negatively associated with trust in the robot, because when less blame is assigned to the actor, more blame is assigned to the robot (Crant & Bateman, 1993). As a result, H2c could be converted to the prediction that two types of trust should be positively associated with non-external control and/or negatively associated with non-personal control.

A linear regression model was tested with external locus of controllability, non-external control, instability, and non-personal control as independent variables and competence-based trust as the dependent variable, and the overall model was significant, R^2 = .09, Adjusted R^2

= .08, $p < .001$, with instability being the only significant predictor, so a simple linear regression was performed in which instability ($\beta = .29$, $p < .001$) was entered as the only explanatory variable, $R^2 = .08$, Adjusted $R^2 = .08$, $p < .001$. This means perceived instability of causality was positively associated with competence-based trust. Hence, H2b was supported for competence-based trust. Another regression model was built with integrity-based trust as a dependent variable, and the model was significant, $R^2 = .10$, Adjusted $R^2 = .10$, $p < .001$, in which instability and non-personal control were significant predictors. Hence, the regression model was rebuilt with the two independent variables, ($\beta = .28$, $p < .001$) and non-personal control ($\beta =$ - .05, $p < .001$), $R^2 = .09$, Adjusted $R^2 = .08$, $p < .001$. This means perceived instability of causality was a positive indicator of integrity-based trust, while nonpersonal control was a negative indicator of integrity-based trust. Therefore, H2b and H2c were supported for integrity-based trust, with H2a failing to gain evidence from the data for either trust type, so overall H2 was partially supported.

H3 suggested apology with both-internal and external-attribution apology would intrigue more (a) unstable and (b) uncontrollable attributions after error occurrence than no repair, so different failure conditions were collapsed into three categories (i.e., internal-attribution apology, external-attribution apology, and no repair). Under the structure of CDSII, H3b could be translated to say that apology should be positively associated with non-external control or negatively associated with non-personal control. For H3a, the assumption of homogeneity was met, $F(2, 503) = 0.83$, $p = .44$, and the group differences in the ANCOVA model with propensity to trust ($p < .01$) and entity beliefs ($p < .05$) as significant covariates and instability as a dependent variable turned out to be insignificant, $F(2, 501) = 0.21$, $p = .81$, partial $\eta^2 = .001$.

For H3b, another two ANCOVA tests were first performed first with non-external control and then non-personal control as the dependent variables. Because both propensity to trust (p = .12) and entity beliefs (p = .60) were insignificant covariates in the first ANCOVA model, an ANOVA model with non-external control as the dependent variable was tested instead. The results of Levene's test were insignificant, $F(2, 503) = 1.57$, $p = .21$, and the group differences turned out to be insignificant for non-external control, $F(2, 503) = 1.52$, $p = .22$, partial $\eta^2 = .01$. In the ANCOVA model with non-personal control as the dependent variable, propensity to trust (p = .12) and entity beliefs (p = .47) were not significant covariates, so another ANOVA model was tested after excluding two covariates. The group effects on non-personal control was insignificant, $F(2, 503) = 0.73$, $p = .48$, partial $\eta^2 = .003$, with the assumption of homogeneity met, $F(2, 503) = 1.15$, $p = .32$. In a word, H3 was rejected.

It was proposed by H4 that denial would elicit more external locus of causality than no repair, so conditions across different failure types that implemented denial and no repair were respectively combined. Since propensity to trust was not a significant covariate, $p = .60$, the ANCOVA test was rerun after removing it. The Levene's test was insignificant at the .01 level, $F(1, 347) = 5.69$, $p = .02$, and the results indicated the group difference was not significant for locus of causality, $F(1, 346) = 0.07$, $p = .79$, partial $\eta^2 = .0002$, so H4 was also rejected.

Results from H5 to H7

H5, H6, and H7 proposed testing the effects of repair methods on competence- and integrity-based trust. H5 predicted internal-attribution apology would be more effective than external-attribution apology, which was examined with two ANCOVAs. In the first model in which competence-based trust was regarded as dependent variables while two conditions of apology were entered as independent variables, entity beliefs (p = .21) were an insignificant covariate and were therefore excluded from the revised model. The result of Levene's test suggested error variances were equal across groups, $F(1, 337) = 1.35$, $p = .25$. Further analysis suggested the group differences yielded insignificant effects on competence-based trust, $F(1, 336) = 0.27$, $p = .60$, partial $\eta^2 = .001$, with propensity to trust being a significant covariate, $F(1, 336) = 10.12$, $p < .01$, partial $\eta^2 = .03$. The mean difference between internal-attribution apology ($M = 7.69$, $SD = 0.13$) and external-attribution apology ($M = 7.60$, $SD = 0.13$), 0.10, 95% CI [-0.27, 0.46], was statistically insignificant, which means internal-attribution apology did not elicit better repair outcomes for competence-based trust than external-apology attribution, so H5 was not supported for competence-based trust.

For integrity-based trust, the assumption of homogeneity of between-group variances was met, F (1, 337) = 0.28, $p = .59$, and the ANCOVA results indicated the main effects were insignificant, F (1, 335) = 2.85, $p = .09$, partial $\eta^2 = .008$, with propensity to trust, $F(1, 335) = 11.81$, $p < .01$, partial $\eta^2 = .03$, and entity beliefs, $F(1, 335) = 2.85$, $p < .01$, partial $\eta^2 = .03$, being two significant covariates. The mean difference for integrity-based trust between internal-attribution apology ($M = 4.71$, $SD = 0.07$) and external-attribution apology ($M = 4.53$, $SD = 0.07$), 0.18, 95% CI [-0.03, 0.39], was insignificant. The findings showed that internal-attribution apology did not generate better repair outcomes than external-attribution apology for integrity-based trust. To sum up, H5 was not supported.

The data lent some support to H6, which predicted external-attribution apology would outperform denial in trust repair. Under equal error variances ($F_{\text{competence-based trust}}$ (1, 343) = 1.76, $p = .19$; $F_{\text{integrity-based trust}}$ (1, 343) = 0.45, $p = .51$) and covariances (Box's $M = 4.45$, $F(3, 34521096) = 1.47$, $p = .22$), the MANCOVA model was significant testing the differences in post-interaction trust caused by the division between external-attribution apology and denial, after excluding entity beliefs (p = .39) as an insignificant covariate, Wilks' Lambda (λ) = .98, F(2, 341) = 3.75, $p < .05$, partial η^2 = .02, with propensity to trust being a significant covariate, Wilks' Lambda (λ) = .97, F(2, 341) = 6.03, $p < .01$, partial η^2 = .03. The mean difference, 0.34, 95% CI [-0.06, 0.75], between external-attribution apology (M = 7.61, SD = 0.15) and denial (M = 7.27, SD = 0.14), was insignificant for competence-based trust, F(1, 342) = 2.80, p = .10, partial η^2 = .08. However, the mean difference, 0.30, 95% CI [0.08, 0.51], between external-attribution apology (M = 4.54, SD = 0.08) and denial (M = 4.25, SD = 0.08), was significant for integrity-based trust, F(1, 342) = 7.27, $p < .01$, partial η^2 = .02. As a result, H6 was partially supported.

H7 focused on the comparisons between denial and no repair, proposing denial would be more harmful than no repair under competence-based trust violations. Another MANCOVA was performed to test the group differences on post-interaction trust, and entity beliefs turned out to be an insignificant covariate, p = .19, without which the model was reconstructed. Box's test of equality of covariance was insignificant, Box's M = 3.80, F(3, 28329051) = 1.26, p = .29, and Levene's tests indicated the assumption of homogeneity of variance was met ($F_{\text{competence-based trust}}$ (1, 347) = 0.94, p = .33; $F_{\text{integrity-based trust}}$ (1, 347) = 1.53, p = .22). The omnibus effects were significant when two dimensions of trust were examined as a set, Wilks' Lambda (λ) = .96, F(2, 345) = 7.17, $p < .01$, partial η^2 = .04, with propensity to trust being a significant covariate, Wilks' Lambda (λ) = .98, F(2, 345) = 3.13, $p < .05$, partial η^2 = .02. On the univariate level, the mean difference between denial (M = 7.27, SD = 0.14) and no repair (M = 7.56, SD = 0.15), -0.29, 95% CI [-0.70, 0.12], p = .17, was insignificant for competence-based trust, but the mean difference between denial (M = 4.25, SD = 0.07) and no repair (M = 4.64, SD = 0.08), -0.39, 95% CI [-0.60, -0.19], $p < .001$, was significant for integrity-based trust. The results confirmed the proposition that denial performed worse as a trust-repairing strategy than no repair for integrity-based violations. Therefore, H7 was partially supported.

Results of RQ3

RQ3 was concerned with the interaction effects between failure types and repair methods, and a two-way MANCOVA model with propensity to trust and entity beliefs as covariates, in which failure types and repair methods were entered as independent variables, and two types of post-interaction trust were entered as dependent variables. Box's test of equality of covariance was insignificant at the .01 level, Box's $M = 45.61$, $F(33, 938504) = 1.36$, $p = .08$, and both dependent variables met the assumption of homogeneity of variance at the .01 level: $F_{\text{competence-based trust}}(11, 676) = 1.95$, $p = .03$; $F_{\text{integrity-based trust}}(11, 676) = 1.42$, $p = .16$. Propensity to trust (Wilks' Lambda (λ) = .97, $F(2, 673) = 10.67$, $p < .001$, partial $\eta^2 = .03$) and entity beliefs (Wilks' Lambda (λ) = .98, $F(2, 673) = 5.38$, $p < .01$, partial $\eta^2 = .02$) were significant covariates in the model; the interaction effects, however, were insignificant in the multivariate model, Wilks' Lambda (λ) = .98, $F(12, 1346) = 1.34$, $p = .19$, partial $\eta^2 = .01$. According to tests of between-subject effects, this interaction was insignificant for both competence-based trust, $F(6, 674) = 1.20$, $p = .31$, partial $\eta^2 = .01$, and integrity-based trust, $F(6, 674) = 1.07$, $p = .38$, partial $\eta^2 = .01$. Therefore, there was no significant interaction between types and repair methods.

Discussion

The current study examined the effects of different types of technical failures made by a robot in human-robot interactions (HRI) and trust repair strategies on human-to-robot trust. Robots have been playing an increasingly critical role in various aspects of human life, and HRI-based trust actively shapes individuals' relationships with these machines. Drawing on the three types of basic technical errors in computer science, the present study developed a three-fold failure taxonomy (i.e., logic, semantic, and syntax failures); based on previous organizational, interpersonal, and human-machine communication (HMC) literature, this study further explored the failure types' interactions with four trust repair methods (i.e., internal-attribution apology, external-attribution apology, denial, and no repair).

The analysis of covariates first indicated propensity to trust (i.e., general trust in robots) was much more closely connected to trust repair than entity beliefs (i.e., to which extent individuals believe personal traits are fixed)—while the former was found significant in most of the aforementioned tests, the latter was only significant concerning integrity-based trust in the ANCOVA model testing the effects of repair types. The analyses of data partially contradicted Kam's (2009) assertion that entity beliefs should be negatively related to trust repair outcomes, and one possible reason is that people conceptualize robotic entities quite differently from how they conceptualize humans, so the scale measuring implicit beliefs in human-human interactions may not be directly translated to the human-machine relationships, which highlights the necessity of developing a scale specifically dedicated to HMC.

In contrast to Marinaccio et al.'s (2015) propositions that slips in HMC should also be integrity-based violations as they do in human-to-human interactions, the current study first postulated that participants would perceive slips, including both logic and semantic failures, to be competence-based violations in HRI because people conceptualize robots to be of agency with less moral and voluntary actions, compared with human beings. This viewpoint was bolstered by the significant results of H1a and partial support for H1b, which indicated apology with internal attributions outperformed denial under both logic and semantic failures. One possibility as to why internal-attribution apology promoted both trust types under logic failures but was only significantly more effective when repairing integrity-based trust under semantic failures could be that the participants considered detecting partially incorrect responses and identifying internal causes to be more intellectually challenging for the robot, while this was less of the case for completely irrelevant output. Nevertheless, taking the responsibilities for technical failures was assessed more positively under each failure category. Given that internal-attribution apology also repairs trust more effectively than denial for competence-based violations and that denial outperforms internal-attribution apology for integrity-based violations in HMC (Quinn, 2018; Sebo et al., 2019), it could be deducted that logic and semantic failures were both competence-based violations rather than integrity-based violations.

If the insignificant results of H1c were not caused by chance occurrences or lack of statistical power, they do however entail some additional questions concerning syntax failures (i.e., lapses). There existed a possibility that different individuals perceived such failures differently when it came to the relationship between competence and integrity, which canceled out the differences in repair effects, or participants simply reacted to two repair strategies in similar manners under this failure condition. When the robot failed to respond, the explanation might seem more logical that some unknown external forces instead of internal causes disturbed its program operation, compared with the other two failure conditions, so the participants were more trusting after the robot responded with denial.

The findings of RQ1 further revealed the three types of failures did not have distinguishable effects on post-interaction evaluations (i.e., perceived competence, integrity, severity of violations, and post-interaction trust), when no repair was implemented. Therefore, the main effects of repair methods and the interaction effects with failure types were the keys for decoding the trust repair outcomes. The findings also added to the research findings of miss- vs. false-prone errors, consistent with Madhavan et al.’s (2006) and Rovira and Parasuraman’s

(2010) conclusions that both types of errors are equally destructive. As it is noted by Hoff and Bashir (2015), a major problem of previous studies on HAI miss vs. false is that two types of errors entail different future risks, which might affect individuals' evaluations of the system: a false alarm might just be disturbing, but a miss might lead to fatal outcomes. Therefore, the direct violation outcomes in the present study were controlled in a way that the human interactants already possessed the access to all information and would immediately point out that NAO failed to provide the correct information after each failure occurrence, which uniformed the direct violation outcomes of each failure type. Under such circumstances, logic, semantic, and syntax did not significantly differ in violation magnitude without trust repair. This suggests controlling for error outcomes of different error types could be helpful for addressing some gaps emerging in the extant error/failure research.

The main effects of failure types were insignificant for competence-based trust and relatively weak for integrity-based trust, and the negative impacts of logic failures were potentially the greatest overall out of three types of failures. Compared to the other two types of failure, logic failures generally presented more accuracy and correctness in the output content, as partially precise responses. One of the explanations of why logic failures were the most detrimental for integrity-based trust is that they appeared harder to catch, which might have raised more doubt for the robot's deliberate deceptions. It was also possible that the partial correctness raised participants' expectations of the robot's performance, so they felt fooled and disappointed after figuring out failure occurrences.

According to the fundamental associations between blame attributions and trust, the current study hypothesized external, unstable, and uncontrollable causal attributions would lead to higher levels of trust. The test results showed the significant associations between non-personal control and integrity-based trust as well as the ones between instability and two types of trust. It was interesting non-personal control was only a significant predictor for integrity-based trust but not for competence-based trust, which indicated participants did not closely connect controllability in HRI blame attributions with robotic intelligence. Noticeably, non-personal control (i.e., the perception that the failure cause is not controlled by the human interactant) turned out to be more reflective of uncontrollability in the context of HRI, as opposed to non-external control (i.e., the perception that the failure cause is not controlled by the NAO robot), which might have resulted from the ontological difference that humans perceive less agency in robots in HRI than they do in humans. For the total rejection of H2a, one possibility of why external locus was not a significant predictor of post-interaction trust might have been that individuals deemed human internal qualities to be less relevant in human-robot interactions because the interactions were considered more impersonal.

Based on the deductions from Baker et al. (2018) and Kim et al. (2004), H3 predicted two types of apology would repair trust through the increase of perceived instability and uncontrollability, whereas H4 proposed denial would boost attributions of external locus. In contrast to the initial expectations, neither of the hypotheses were supported by the analysis. It was probable that some unidentified interactions amongst failure, repair types, and blame attributions canceled out the differences on these dimensions of causality, if it were not for the problem of insufficient statistical power under very small effects, or oral accounts were not powerful enough to alert the participants' blame attributions on a conscious level, corresponding with Schweitzer et al.'s (2006) opinion that a single apology without actual behavioral improvements might not be powerful enough in changing people's opinions.

Contradicted with the previous findings from Kim et al. (2006), the present study found the rule in human-human communication might not be applicable under the context of HRI that internal-attribution apology can repair competence-based trust violations more effectively than external-attribution apology, exploring the effects of two apology types on two subdimensions of trust (i.e., competence- and integrity-based trust). Prior studies already pointed out the potential risks lying underneath external-attribution apology are that it might negatively impact perceived integrity in interpersonal interactions (Schlenker et al., 2001; Tomlinson et al., 2004), and the insignificant test results indicated the negativity of external attributions in apology might be dissimilar or much smaller under the context of HRI. It was possible that the participants conceptualized robots with less moral agency, so they were less inclined to surmise the NAO robot intended to lie when it gave external-attribution apology.

The partial support of H6, on the other hand, gives some more insights into the nature of external apology and denial. Based on hierarchically restrictive schemas (Reeder & Brewer, 1979), Kim et al. (2004) argued that apology works better than denial for competence-based trust violations, because people attach more importance to positive information than negative information in this kind of situation. Following this logic, external-attribution apology should have been more trust-gaining than denial since the NAO robot expressed remorse and made promises for the future, which was confirmed by the test results. Participants might have perceived external-attribution apology to be more honest than denial, which completely shirked the blame, since external-attribution apology afforded partial responsibilities in addition to expressed remorse and given promises, even though addressing the former was not necessarily perceived as more intelligent than delivering the latter. Additionally, it was proved that denial in general repaired trust much worse than no repair. This further suggested the detrimental power of eluding the blame for competence-based violations—it is not always the case that taking action is better than not taking action, when the action is deemed inappropriate and unpleasant in HRI, such as a robot denying mistakes or blaming someone else for its own mistake.

Finally, the insignificant interactions between failure and repair types further emphasized the similarity amongst three failure types as competence-based violations, if not type II error. This implicated the principles that internal- and external-attribution apology both worked better in trust repair than denial in the context of HRI, no matter which type of competence-based violations are there.

Theoretical and Practical Implications

The present study developed a new categorization of technical failures that is message-based and recipient-oriented from a communication-centered perspective, contributing to the extant research in robotic failures and errors. Because of its cross-contextual applicability, this categorization can be easily applied under other HMC contexts, such as human-automation interactions, human-agent interactions, and human-computer interactions. The findings also denoted that people may perceive the division between competence- and integrity-based violations differently in HMC, compared to how they process information in human-to-human communication. The systematic investigation into four different repair methods filled in some gaps of prior literature, such as the comparisons between external-attribution apology and denial which previous studies did not examine. The findings also suggested the redirection of attributions by different repair methods might be more complex than expected, contributing to the extant trust repair research.

Pragmatically speaking, the research underlined how the technological ability to detect and respond to failures could enhance user trust. This study could benefit technical designs of robots, given the prevalence of these failures in robots and many other kinds of technologies. It provides a user-oriented perspective for understanding the impacts of common technical failures: instead of approaching these failures from error mechanisms, it might be more helpful to inquire what end users perceive the causes to be and implement repair strategies accordingly. The findings over failure types and repair methods in the present study could help designers identify the optimum repair strategies under each failure type, promoting both short-term and long-term human-to-machine trust: altogether, apology is more helpful than denial, no matter whether the failure was logical, semantic, or syntactic. Thus, it is recommended to program such apologetic speech acts in robots when responding to humans pointing out any mistake they make and unsatisfied with their performances.

Limitations

Since previous studies showed demographic characteristics, such as age, culture, gender, and occupations, have noticeable impacts on HRI-base trust patterns (Hoff & Bashir, 2015). One major limitation of the present study is that the analysis results from college student sample will not be generalizable to other social groups. Another constraint resulted from the experimental designs: the between-within subject designs might have introduced some undesirable bias in the statistical tests. The in-lab design of one-time contact also limited generalizability of the results concerning the long-term impacts of technical failures on actual human-robot relationships. Moreover, the design of inducing HRI scenarios through online videos might have also posited some methodological limitations, given participants might not have been as involved as they would when they are presented with interventions of greater interactivity and social presence (e.g., interactive videos, live interactions; Xu et al., 2015), which could have affected their blame attributions: for example, they might have made less internal attributions as they identified less with the focal person as an observer.

Directions for Future Research

The present study also suggested some possibilities for future investigation. First and foremost, future research could look into how other demographic factors influence failure perceptions and trust repair outcomes. Take the age groups for example, they might possess distinct response patterns to robotic failures because of greater trust inertia. It will also be interesting to explore cultural differences in trust repair preferences, considering diversified cultural norms and distinct usage patterns of robots in different societies. Another indication is that future research could investigate other failure taxonomies, such as non-critical, repairable, vs. terminal failures (Carlson & Murphy, 2005) and technical vs. decision-level failures (Flook et al., 2019), or more repair strategies, including empathizing, emotionally regulating, recognizing, anthropomorphizing, trumping, downgrading, and gaslighting (de Visser et al., 2018). Considering the participants' connections with the robot in the present study were completely experiment-based, it will be valuable to investigate how people deal with different failures and repair strategies with technologies they actually use outside of labs, such as computers and personal digital assistants (e.g., Apple's Siri, the Google Assistant, Alexa), considering the dynamic of trust is rather complex. The other alternative is to extend the one-time contact into multiple-time interactions in order to observe both short- and long-term impacts of robotic failures on HRI-based trust. Last but not least, future research can also study how differences in the abilities to cope with technical failures may contribute to digital inequality.

Reference

Aquino, K., Tripp, T. M., & Bies, R. J. (2001). How employees respond to personal offense: The effects of blame attribution, victim status, and offender status on revenge and reconciliation in the workplace. *Journal of Applied Psychology*, *86*(1), 52–59.

Bachmann, R., Gillespie, N., & Priem, R. (2015). Repairing trust in organizations and institutions: Toward a conceptual framework. *Organization Studies*, *36*(9), 1123–1142.

Baker, A. L., Phillips, E. K., Ullman, D., & Keebler, J. R. (2018). Toward an understanding of trust repair in human-robot interaction. *ACM Transactions on Interactive Intelligent Systems*, *8*(4), 1–30.

Banks, J. (2019). A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior*, *90*, 363–371.

Bansal, G., & Zahedi, F. M. (2015). Trust violation and repair: The information privacy perspective. *Decision Support Systems*, *71*, 62–77.

Beldad, A. D., Van Laar, E., & Hegner, S. M. (2018). Should the shady steal thunder? The effects of crisis communication timing, pre-crisis reputation valence, and crisis type on post-crisis organizational trust and purchase intention. *Journal of Contingencies and Crisis Management*, *26*(1), 150–163.

Bies, R. J., & Tripp, T. M. (1996). Beyond distrust: Getting even and the need for revenge, In R. M. Kramer & T. R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research* (pp. 246–260). Sage Publications.

Bradfield, M., & Aquino, K. (1999). The effects of blame attributions and offender likableness on forgiveness and revenge in the workplace. *Journal of management*, *25*(5), 607–631.

Brooks, D. 2017. *A human-centric approach to autonomous robot failures* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Database. (UMI No. 10643702)

Carlson, J., & Murphy, R. R. (2005). How UGVs physically fail in the field. *IEEE Transactions on Robotics*, *21*(3), 423–437.

Chamani, F., & Zareipur, P. (2010). A cross-cultural study of apologies in British English and Persian. *Concentric: Studies in Linguistics*, *36*(1), 133–153.

Chancey, E. T., Bliss, J. P., Liechty, M., & Proaps, A. B. (2015, September). False alarms vs. misses: Subjective trust as a mediator between reliability and alarm reaction measures. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 647–651).

Chiu, C. Y., Dweck, C. S., Tong, J. Y., & Fu, J. H. (1997). Implicit theories and conceptions of morality. *Journal of Personality and Social Psychology*, *73*(5), 923–940.

Coeckelbergh, M. (2012). Can we trust robots?. *Ethics and Information Technology*, *14*(1), 53–60.

Cohen, A. D., & Olshtain, E. (1985). Comparing apologies across languages. In K. R. Jankowsky (Ed.), *Scientific and humanistic dimensions of language* (pp.175–184). John Benjamins.

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, *92*, 909–927.

Correia, F., Guerra, C., Mascarenhas, S., Melo, F. S., & Paiva, A. (2018, July). Exploring the impact of fault justification in human-robot trust. *Proceedings of the 17th International Conference on Autonomous Agents and Multi Agent Systems*, 507–513.

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human-Computer Studies*, *58*(6), 737–758.

Crant, J. M., & Bateman, T. S. (1993). Assignment of credit and blame for performance outcomes. *Academy of Management Journal*, *36*(1), 7–27.

Davenport, R. B., & Bustamante, E. A. (2010, October). Effects of false alarm vs. miss-prone automation and likelihood alarm technology on trust, reliance, and compliance in a miss-prone task. *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting*, 1513–1517.

Davenport, R. B., & Bustamante, E. A. (2010). Effects of false-alarm vs. miss-prone automation and likelihood alarm technology on trust, reliance, and compliance in a miss-prone task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(19), 1513–1517.

De Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., & Parasuraman, R. (2012, September). The world is not enough: Trust in cognitive agents. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1), 263–267.

De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349.

De Visser, E. J., Pak, R., & Neerincx, M. A. (2017, March). Trust development and repair in human-robot teams. *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 103–104.

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From “automation” to “autonomy”: The importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409–1427.

De Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics*, *12*(2), 459–478.

De Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, *58*(6), 719–735.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. *8th ACM/IEEE International Conference on Human-Robot Interaction*, 251–258.

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfield, A., & Yanco, H. (2012). Effects of changing reliability on trust of robot systems. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, *7*, 73–80.

Desai, M., Stubbs, K., Steinfeld, A., & Yanco, H. (2009, April). *Creating trustworthy robots: Lessons and inspirations from automated systems* [Conference presentation]. The AISB Convention: New Frontiers in Human-Robot Interaction, Edinburgh, Scotland.

DeVellis, R. F. (2016). Scale development: Theory and applications. *Sage publications*.

Dirks, K. T., Lewicki, R. J., & Zaheer, A. (2009). Reparing relationships within and between organizations: Building a conceptual foundation. *Academy of Management Review*, *34*(1), 68–84.

Dixon, S. R. (2006). *Imperfect diagnostic automation: How adjusting bias and saliency affects operator trust* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Database. UMI No.3242835

Dixon, S. R., & Wickens, C. D. (2003). *Imperfect automation in unmanned aerial vehicle flight control* [Report No. AHFD-03-17/MAAD-03-2]. Aviation Human Factors Division.

Dixon, S. R., & Wickens, C. D. (2004). Reliability in automated aids for unmanned aerial vehicle flight control: Evaluating a model of automation dependence in high workload. [Report No. AHFD-04-05/MAAD-04-1]. Aviation Human Factors Division.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(3), 474–486.

Doney, P. M., Cannon, J. P., & Mullen, M. R. (1998). Understanding the influence of national culture on the development of trust. *The Academy of Management Review,* 23(3), 601–620.

Dweck, C. S., Chiu, C. Y., & Hong, Y. Y. (1995). Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological Inquiry*, *6*(4), 267–285.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-computer Studies*, *58*(6), 697–718.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *44,* 79–97.

Eberl, P., Geiger, D., & Aßländer, M. S. (2015). Repairing trust in an organization after integrity violations: The ambivalence of organizational rule adjustments. *Organization Studies*, *36*(9), 1205–1235.

Evers, V., Maldonado, H., Brodecki, T., & Hinds, P. (2008, March). Relational vs. group self-construal: untangling the role of national culture in HRI. *3rd ACM/IEEE International Conference on Human-Robot Interaction*, 255-262).

Ferrell, C. (1994). Failure recognition and fault tolerance of an autonomous robot. *Adaptive behavior*, *2*(4), 375–398.

Ferrin, D. L., Kim, P. H., Cooper, C. D., & Dirks, K. T. (2007). Silence speaks volumes: The effectiveness of reticence in comparison to apology and denial for responding to integrity- and competence-based trust violations. *Journal of Applied Psychology*, *92*(4), 893–908.

Flook, R., Shrinah, A., Wijnen, L., Eder, K., Melhuish, C., & Lemaignan, S. (2019). On the impact of different types of errors on trust in human-robot interaction: Are laboratory-based HRI experiments trustworthy?. *Interaction Studies*, *20*(3), 455-486.

Fuoli, M., van de Weijer, J., & Paradis, C. (2017). Denial outperforms apology in repairing organizational trust despite strong evidence of guilt. *Public Relations Review*, *43*(4), 645–660.

Garcia, C. (1989). Apologizing in English: Politeness strategies used by native and non-native speakers. *Multilingua-Journal of Cross-Cultural and Interlanguage Communication*, *8*(1), 3–20.

Geels-Blair, K., Rice, S., & Schwark, J. (2013). Using system-wide trust theory to reveal the contagion effects of automation false alarms and misses on compliance and reliance in a simulated aviation task. *The International Journal of Aviation Psychology*, *23*(3), 245–266.

Gervey, B. M., Chiu, C. Y., Hong, Y. Y., & Dweck, C. S. (1999). Differential use of person information in decisions about guilt versus innocence: The role of implicit theories. *Personality and Social Psychology Bulletin*, *25*(1), 17–27.

Gheorghiu, M. A., Vignoles, V. L., & Smith, P. B. (2009). Beyond the United States and Japan: Testing Yamagishi's emancipation theory of trust across 31 nations. *Social Psychology Quarterly*, *72*(4), 365–383.

Gillespie, N., & Dietz, G. (2009). Trust repair after an organization-level failure. *Academy of Management Review*, *34*(1), 127–145.

Gillespie, N., Dietz, G., & Lockey, S. (2014). Organizational reintegration and trust repair after an integrity violation: A case study. Business Ethics Quarterly, *24*(3), 371–410.

Giuliani, M., Mirnig, N., Stollnberger, G., Stadler, S., Buchner, R., & Tscheligi, M. (2015). Systematic analysis of video data from different human-robot interaction studies: A categorization of social signals during error situations. *Psychol*., *6*, 931–931.

Goles, T., Rao, S. V., Lee, S., & Warren, J. (2009). Trust violation in electronic commerce: Customer concerns and reactions. *Journal of Computer Information Systems*, *49*(4), 1–9.

Gray, H. M, Gray, K, & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124.

Groom, V., Chen, J., Johnson, T., Kara, F. A., & Nass, C. (2010, March). Critic, compatriot, or chump? Responses to robot blame attribution. *2010 5th ACM/IEEE International Conference on Human-Robot Interactions*, 211-217.

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New Media & Society*, 22(1), 70–86.

Guznov, S., Lyons, J., Nelson, A., & Woolley, M. (2016). The effects of automation error types on operators' trust and reliance. In International Conference on Virtual, Augmented and Mixed Reality, 116–124.

Haesevoets, T., Chris, R. F., & Alain, V. H. (2015). Is trust for sale? the effectiveness of financial compensation for repairing competence- versus integrity-based trust violations. *PLoS One*, *10*(12). Retrieved from doi:http://dx.doi.org.ezproxy.lib.ou.edu/10.1371/journal.pone.0145952

Hall, E. (1977). *Beyond culture*. Anchor.

Hancock, P., Billings, D., Schaefer, K., Chen, J., De Visser, E., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *53*(5), 517–527.

Heider. F (1958). The psychology of interpersonal relations. New York: Wiley.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *57*(3), 407–434.

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems*, *28*(1), 84–88.

Honig, S., & Oron-Gilad, T. (2018). Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in Psychology*, *9*, 861–861.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: A multidisciplinary journal*, *6*(1), 1–55.

Huerta, E., Glandon, T., & Petrides, Y. (2012). Framing, decision aid systems, and culture: Exploring influences on fraud investigations. *International Journal of Accounting Information Systems*, *13*, 316–333.

Huff, L., & Kelley, L. (2005). Is collectivism a liability? The impact of culture on organizational trust and customer orientation: A seven-nation study. *Journal of Business Research, 58*(1), 96–102.

Iio, T., Yoshikawa, Y., Chiba, M., Asami, T., Isoda, Y., & Ishiguro, H. (2020). Twin-robot dialogue system with robustness against speech recognition failure in human-robot dialogue with elderly people. *Applied Sciences*, *10*(4), 1522–1522.

Janowicz-Panjaitan, M., & Krishnan, R. (2009). Measures for dealing with competence and integrity violations of interorganizational trust at the corporate and operating levels of organizational hierarchy. *Journal of Management Studies*, *46*(2), 245–268.

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019, July). The measurement of the propensity to trust automation. *International Conference on Human-Computer Interaction*, 476–489. Springer, Cham.

Johnson, J. D., Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004, September). Type of automation failure: The effects on trust and reliance in automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(18), 2163–2167.

Jung, E. H. (1999). The acquisition of communicative competence in a second language. *Journal of Pan-Pacific Association of Applied Linguistics*, *3*, 13–37.

Kam, T. K. (2009, June). *Implicit theories and the trust repair process* [Conference presentation]. *22nd Annual IACM Conference*, Kyoto, Japan.

Kaniarasu, P., & Steinfeld, A. M. (2014, August). Effects of blame on trust in human robot interaction. *23rd IEEE International Symposium on Robot and Human Interactive Communication*, 850–855.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015). Comparing robot embodiments in a guided discovery learning interaction with children. *International Journal of Social Robotics*, *7*(2), 293–308.

Kim, P. H., Cooper, C. D., Dirks, K. T., & Ferrin, D. L. (2013). Repairing trust with individuals vs. groups. *Organizational Behavior and Human Decision Processes*, *120*(1), 1–14.

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. *Academy of Management Review*, *34*(3), 401–422.

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation. *Organizational behavior and human decision processes*, *99*(1), 49–65.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of Applied Psychology*, *89*(1), 104–118.

Kohn, S. C., Quinn, D., Pak, R., de Visser, E. J., & Shaw, T. H. (2018, September). Trust repair strategies with self-driving vehicles: An exploratory study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 1108–1112.

Kontogiorgos, D., van Waveren, S., Wallberg, O., Pereira, A., Leite, I., & Gustafson, J. (2020, April). Embodiment effects in interactions with failing robots. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology, 50*, 569–598.

Kramer, R. M., & Lewicki, R. J. (2010). Repairing and enhancing trust: Approaches to reducing organizational trust deficits. *Academy of Management Annals*, *4*(1), 245–277.

Krosgaard, M. A., Brodt, S. E., & Whitener, E. M. (2002). Trust in the face of conflict: The role of managerial trustworthy behavior and organizational context. *Journal of Applied Psychology, 87*(2), 312–319.

Laprie, J. C. (1985, June). Dependable computing and fault-tolerance. *The International Symposium on Fault-Tolerant Computing*, *15*, 2–11.

Laprie, J. C. (1995, June). Dependable computing and fault tolerance: concepts and terminology. *International Symposium on Fault-Tolerant Computing*, *25*, 2–11.

Lee, John D, & See, Katrina A. (2004). trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80.

Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., & Rybski, P. (2010, March). Gracefully mitigating breakdowns in robotic services. *5th ACM/IEEE International Conference on Human-Robot Interaction*, 203–210.

Levinthal, B. R., & Wickens, C. D. (2006, October). Management of multiple UAVs with imperfect automation. *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*, 1941–1944.

Lewicki, R. J., & Brinsfield, C. (2017). Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior, 4*, 287–313.

Lewicki, R. J., & Tomlinson, E. C. (2003, June). The effects of reputation and post violation communication on trust and distrust. *16th Annual International Association for Conflict Management Conference*, Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=400941

Lucas, G. M., Boberg, J., Traum, D., Artstein, R., Gratch, J., Gainer, A., Johnson, E., Leuski, A., & Nakano, M. (2018). Getting to know each other: The role of social dialogue in recovery from errors in social robots. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 344–351.

Madhavan, D., & Wiegmann, D. A. (2005). Effects of information source, pedigree, and reliability on operators utilization of diagnostic advice. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 49*(3), 487–491.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *48*, 241–256.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016, March). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. *11th ACM/IEEE International Conference on Human-Robot Interaction,* 125–132.

Marinaccio, K., Kohn, S., Parasuraman, R., & De Visser, E. J. (2015, June*).* A framework for rebuilding trust in social automation across health-care domains. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care, 4*(1), 201–205.

Mayer, R. C, Davis, J. H., & Schoorman, F. D. 1995. An integrative model of organizational trust. *Academy of Management Review*, 20, 709–734.

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology, 84*(1), 123–136.

McAuley, E., Duncan, T. E., & Russell, D. W. (1992). Measuring causal attributions: The revised causal dimension scale (CDSII). *Personality and Social Psychology Bulletin*, *18*(5), 566–573.

McCall, D., & Kölling, M. (2014 October). Meaningful categorisation of novice programmer errors". *Proceedings of 2014 IEEE Frontiers in Education Conference*, 1–8.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(2), 194–210.

Meyerson, D., Weick, K. E., & Kramer, R. M. (1996). Swift trust and temporary groups. In R. M. Kramer & T. R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research* (pp. 166–195). Sage Publications.

Mir, M. (1992). Do we all apologize the same? An empirical study on the act of apologizing by Spanish speakers learning English. *Pragmatics and Language Learning*, *3*, 1–19.

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017, May). To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI*. Retrieved from https://doi.org/10.3389/frobt.2017.00021

Nass, C. I., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103.

Osborne, J. W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *52*(3), 381–410.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *39*(2), 230–253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, & Cybernetics*, *30*, 286–297.

Park, E., Kim, K. J., & Del Pobil, A. P. (2011, November). The effects of a robot instructor's positive vs. negative feedbacks on attraction and acceptance towards the robot in classroom. *International Conference on Social Robotics*, 135–141.

Pelikan, H. R., & Broth, M. (2016, May). Why that NAO? How humans adapt to a conventional humanoid robot in taking turns-at-talk. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4921–4932.

Pino, O., Palestra, G., Trevino, R., & De Carolis, B. (2020). The humanoid robot NAO as trainer in a memory program for elderly people with mild cognitive impairment. *International Journal of Social Robotics*, *12*(1), 21–33.

Poppo, L., & Schepker, D. J. (2010). Repairing public trust in organizations. *Corporate Reputation Review*, *13*(2), 124–141.

Powers, A., Kiesler, S., Fussell, S., & Torrey, C. (2007, March). Comparing a computer agent with a humanoid robot. *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, 145–152.

Prendinger, H., & Ishizuka, M. (2004). Introducing the cast for social computing: Life-like characters. In H. Prendinger & M. Ishizuka (Eds.), *Life-like characters: Tools, affective functions, and applications* (pp. 3–16). Springer.

Quinn, D. B. (2018). *Exploring the efficacy of social trust repair in human automation interactions* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Database. (UMI No.10812087).

Rao, V. S., & Lee, S. J. (2007). Responses to trust violation: A theoretical framework. *The Journal of Computer Information Systems*, *48*(1), 76–87.

Reason, J. (1990). *Human error*. Cambridge University Press.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*(1), 61–79.

Reeves, B., & Nass, C. I. (1996). The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press.

Robert, L. P., Denis, A. R., & Hung, Y. T. C. (2009). Individual swift trust and knowledge-based trust in face-to-face and virtual team members. *Journal of Management Information Systems*, *26*(2), 241–279.

Robinette, P., Howard, A. M., & Wagner, A. R. (2015, October). Timing is key for robot trust repair. In A. Tapus, E. André, J. Martin, F. Ferland, & M. Ammi (Eds.), *Social Robotics* (pp. 574–583). Springer.

Ross, R., Collier, R., & O'Hare, G. M. (2004, July). Demonstrating social error recovery with agent factory. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, *3*, 1424–1425.

Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2017, November). How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario. In A. Kheddar, E. Yoshida, S. S. Ge., K. Suzuki, J. Cabibihan, F. Eyssel, H, He. (Eds.), *Social Robotics* (pp. 42–52). Springer.

Rovira, E., & Parasuraman, R. (2010). Transitioning to future air traffic management: Effects of imperfect automation on controller attention and performance. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *52*(3), 411–425.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. *10th ACM/IEEE International Conference on Human-Robot Interaction*, 1–8.

Salem, M., Ziadee, M., & Sakr, M. (2014, March). Marhaba, how may I help you? Effects of politeness and culture on robot acceptance and anthropomorphization. *9th ACM/IEEE International Conference on Human-Robot Interaction*, 74–81.

Sanchez, J. (2006). *Factors that affect trust and reliance on an automated aid* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Database. (UMI No. 3212291).

Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004, September). Reliability and age-related effects on trust and reliance of a decision support aid. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*, 586–589.

Sanchez, J., Rogers, W. A., Fisk, A. D., & Rovira, E. (2014). Understanding reliance on automation: Effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science*, *15*(2), 134–160.

Sanders, T., Kaplan, A., Koch, R., Schwartz, M., & Hancock, P. A. (2019). The relationship between trust and use choice in human-robot interaction. *Human Factors: The Journal of Human Factors and Ergonomics Society, 61*(4), 614–626.

Schaefer, K. (2013). *The Perception and Measurement of Human-Robot Trust* (Doctoral dissertation). Retrieved from https://stars.library.ucf.edu/etd/2688/

Schaefer, K. E., Sanders, T. L., Yordon, R. E., Billings, D. R., & Hancock, P. A. (2012, September). Classification of robot form: Factors predicting perceived trustworthiness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 1548–1552.

Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes, 101*(1), 1–19.

Sebo, S. S., Krishnamurthi, P., & Scassellati, B. (2019, March). “I don't believe you”: Investigating the effects of robot trust violation and repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction*, 57–65.

Sebo, S. S., Traeger, M., Jung, M., & Scassellati, B. (2018, February). The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 178–186.

Shamsuddin, S., Yussof, H., Ismail, L., Hanapiah, F. A., Mohamed, S., Piah, H. A., & Zahari, N. I. (2012, March). Initial response of autistic children in human-robot interaction therapy with humanoid robot NAO. *IEEE 8th International Colloquium on Signal Processing and its Applications*, 188–193.

Shaw, J. C., Wild, E., & Colquitt, J. A. (2003). To justify or excuse?: A meta-analytic review of the effects of explanations. *Journal of Applied Psychology*, *88*(3), 444–458.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, *31*(2), 47–53.

Sitkin, S. B., & Roth, N. L. (1993). Explaining the limited effectiveness of legalistic “remedies” for trust/distrust. *Organization Science*, *4*(3), 367–392.

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-computer Studies*, *52(*4), 701–717.

Smith, C. A., & Lazarus, R. S. (1993). Appraisal components, core relational themes, and emotions. In N. H. Frijda (Ed.), *Appraisal and Beyond: The issue of cognitive determinants of emotion* (pp. 233–269*).* Erlbaum.

SoftBank Robotics. NAO. Retrieved on 2020, November 14th from https://www.softbankrobotics.com/emea/en/nao

Steinbauer, G. (2012, June). A survey about faults of robots used in robocup. *Robot Soccer World Cup*, 344–355.

Stouten, J., De Cremer, D., & Van Dijk, E. (2006). Violating equality in social dilemmas: Emotional and retributive reactions as a function of trust, attribution, and honesty. *Personality and Social Psychology Bulletin*, *32*(7), 894–906.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed). Pearson.

Takaku, S. (2001). The effects of apology and perspective taking on interpersonal forgiveness: A dissonance-attribution model of interpersonal forgiveness. *The Journal of Social Psychology*, *141*(4), 494–508.

Tomlinson, E. C., Dineen, B. R., & Lewicki, R. J. (2004). The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise. *Journal of Management*, *30*(2), 165–187.

Tomlinson, Edward C, & Mryer, Roger C. (2009). The role of causal attribution dimensions in trust repair. *The Academy of Management Review*, *34*(1), 85–104.

Tzeng, J. Y. (2004). Toward a more civilized design: Studying the effects of computers that apologize. *International Journal of Human-Computer Studies*, *61*(3), 319–345.

Utz, S., Matzat, U., & Snijders, C. (2009). On-line reputation systems: The effects of feedback comments and reactions on building and rebuilding trust in on-line auctions. *International Journal of Electronic Commerce*, *13*(3), 95–118.

Vega, A., Ramírez-Benavides, K., Guerrero, L. A., & López, G. (2019). Evaluating the NAO robot in the role of personal assistant: The effect of gender in robot performance evaluation. *Multidisciplinary Digital Publishing Institute Proceedings*, *31*(1), 20–20.

Wagner, A. R. (2016). *Trust and trustworthiness in human-robot interaction: A formal conceptualization* (Final Report). Georgia Tech Research Institute, Air Force Research Laboratory website: https://apps.dtic.mil/sti/pdfs/AD1011180.pdf

Walfisch, T., Van Dijk, D., & Kark, R. (2013). Do you really expect me to apologize? The impact of status and gender on the effectiveness of an apology in the workplace. *Journal of Applied Social Psychology*, *43*(7), 1446–1458.

Wang, L., Rau, P. L. P., Evers, V., Robinson, B. K., & Hinds, P. (2010, March). When in Rome: the role of culture & context in adherence to robot recommendations. *5th ACM/IEEE International Conference on Human-Robot Interaction*, 359–366. IEEE.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548–573.

Weiner, B. (2008). Reflections on the history of attribution theory and research. *Social Psychology*, *39*(3), 151–156.

Weun, S., Beatty, S. E., & Jones, M. A. (2004). The impact of service failure severity on service recovery evaluations and post-recovery relationships. *The Journal of Services Marketing*, *18*(2), 133–146.

Wijnen, L., Coenen, J., & Grzyb, B. (2017, March). “It's not my fault!”: Investigating the effects of the deceptive behaviour of a humanoid robot. *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 321–322.

Xu, Q., Ng, J., Tan, O., Huang, Z., Tay, B., & Park, T. (2015). Methodological issues in scenario-based evaluation of human–robot interaction. *International Journal of Social Robotics*, *7*(2), 279–291.