

Please quote as: Meywirth, S., Janson, A. & Söllner, M. (2026). When Trust Breaks: A Framework For Analyzing And Understanding Trust Violation And Repair In Sociotechnical Systems. European Conference on Information Systems (ECIS), Milan, Italy.

Association for Information Systems

AIS Electronic Library (AISeL)

ECIS 2026 Proceedings

European Conference on Information Systems
(ECIS)

June 2026

When Trust Breaks: A Framework For Analyzing And Understanding Trust Violation And Repair In Sociotechnical Systems

Sophia Meywirth

University of Kassel, sophia.meywirth@uni-kassel.de

Andreas Janson

Institute of Information Systems and Digital Business, andreas.janson@unisg.ch

Matthias Söllner

University of Kassel, soellner@uni-kassel.de

Follow this and additional works at: <https://aisel.aisnet.org/ecis2026>

Recommended Citation

Meywirth, Sophia; Janson, Andreas; and Söllner, Matthias, "When Trust Breaks: A Framework For Analyzing And Understanding Trust Violation And Repair In Sociotechnical Systems" (2026). *ECIS 2026 Proceedings*. 19.

<https://aisel.aisnet.org/ecis2026/litrev/litrev/19>

This material is brought to you by the European Conference on Information Systems (ECIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ECIS 2026 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

WHEN TRUST BREAKS: A FRAMEWORK FOR ANALYZING AND UNDERSTANDING TRUST VIOLATION AND REPAIR IN SOCIOTECHNICAL SYSTEMS

Completed Research Paper

Sophia Meywirth, University of Kassel, Kassel, Germany, sophia.meywirth@uni-kassel.de

Andreas Janson, University of St. Gallen, St. Gallen, Switzerland, andreas.janson@unisg.ch

Matthias Söllner, University of Kassel, Kassel, Germany, soellner@uni-kassel.de

Abstract

Trust violations in sociotechnical systems are becoming increasingly visible as AI systems take on advisory and decision-making roles in everyday life. When these systems fail, users must interpret what went wrong and decide whether to continue relying on them. Yet IS research has largely focused on trust formation, leaving the processes through which trust is broken, and potentially repaired, under-theorized. Meanwhile, adjacent fields have produced fragmented insights into trust violations and the mechanisms used to repair them. This paper presents a systematic literature synthesis of 62 studies to understand how trust violations and repair are conceptualized when technology plays a meaningful role. We identify four core stages that structure trust breakdown and recovery: the violation incident, the trustor sensemaking, the violator reaction, and the recalibrated trust outcome. We consolidate these insights into a process-oriented framework and guiding questions, outlining implications for human-AI collaboration and future IS research.

Keywords: Trust, Trust Violation, Trust Repair, Human-AI Collaboration, Sociotechnical Systems

1 Introduction

In August 2025, news emerged about a Spanish influencer who broke down in tears after missing her flight to Puerto Rico (Chinman, 2025). She relied on ChatGPT for visa advice, which incorrectly assured her that none was required. However, an ESTA was still required, and she consequently missed her flight. In response to the incident, she said through tears, “That’s what I get for not getting more information,” and added, “I don’t trust that one anymore,” referring to her AI assistant.

This is a clear case of misplaced trust and subsequent loss of trust. Yet it also raises broader questions about how such incidents are interpreted and addressed in sociotechnical systems: How do users make sense of such a violation incident, and under what conditions is it perceived as a trust violation? How is responsibility attributed between the user (“I should have checked”) and the AI system (“I don’t trust it anymore”)? And what kinds of system responses enable trust to be maintained, recalibrated, or repaired?

We understand sociotechnical system as a system in which social and technical components are interdependent and jointly shape system outcomes through their interaction (Emery, 2016). Within such systems, how users interpret a trust violation has immediate consequences for their behavior: they may abandon a system, restrict their reliance on it, or continue using it with adjusted expectations (Lee & See, 2004). A violation also recalibrates trust – sometimes in productive ways, when users recognize system limitations, but sometimes in ways that lead to overreaction and unwarranted distrust (Pak & Rovira, 2024). This makes trust violation just as central as trust formation, since both processes influence how people ultimately use and depend on technology.

While trust has been widely studied in information systems research (e.g. Söllner et al., 2018) much of this work has focused on the formation of initial trust in online environments (e.g. Zahedi & Song, 2008), with comparatively less attention paid to trust violations and their consequences. The few exceptions have concentrated on domains such as e-commerce transactions or privacy and data breaches (Bachura et al., 2022; Choi et al., 2016; Deng et al., 2022). Yet these studies primarily address trust between buyers and sellers or between individuals and organizations, rather than violations that emerge directly in human-technology interaction.

By contrast, other disciplines have begun to study how trust can be violated in interactions with non-human agents. For example, research in human-robot interaction (HRI) has examined situations in which robots violate trust through errors or inappropriate behavior (e.g. Esterwood & Robert, 2023; Maehigashi et al., 2025; Sebo et al., 2020), while work on autonomous vehicles has investigated how accidents or malfunctions affect drivers’ willingness to rely on automation (Liu et al., 2024).

These cases suggest that trust violations become particularly salient when technology operates with a degree of agency and when humans must collaborate with non-human agents. With the rapid diffusion of AI-based information systems, such conditions are now increasingly present in non-embodied contexts. This creates ambiguity in responsibility and repair, making them important contexts for studying how trust violations are interpreted, how responsibility is constructed through trustor sensemaking, and how trust can be potentially restored.

Because research directly addressing trust violations in AI systems is still emerging, and because foundational insights are scattered across various sociotechnical contexts, there is a need for a systematic synthesis of what is currently known about trust violations and repair when technology is involved. Such a synthesis provides a foundational framework for understanding how these processes may operate in AI systems specifically and helps identify where AI introduces novel challenges.

In consequence, we investigate the following research question (RQ):

RQ: How have trust violations and their repair been conceptualized across sociotechnical systems, and what insights do these accounts provide into the relationship between trustor sensemaking and violator reactions?

To address this need, we conduct a literature review following established guidelines (vom Brocke et al., 2015; Webster & Watson, 2002). We review studies on trust violations and repair mechanisms across sociotechnical contexts in which technology plays a meaningful role in the violation incident. Based on a concept matrix, we examine how violations are conceptualized, how users’ sensemaking is described,

and what forms of responses or repair strategies have been studied. Based on this synthesis, we develop a framework that clarifies the relationship between victim sensemaking and violator reactions and outline implications for research on trust violations in AI systems.

2 Theoretical Background

2.1 Trust

In this work, we adopt the definition of trust proposed by Mayer et al. (1995, p. 712): “Trust is the willingness of a party [the trustor] to be vulnerable to the actions of another party [the trustee] based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.” In this sense, vulnerability is fundamental to trust, understood as a willingness to be exposed to the potential negative consequences of another party’s actions, with actual risk-taking emerging only when this willingness is acted upon (Mayer et al., 1995).

Building on this definition, IS research emphasizes that trust is an evolving process shaped by the trustor’s propensity, the context, and the trustee’s characteristics (Mayer et al., 1995). In this view, trust is reflected in the trustor’s trusting beliefs, that is, their perceptions of the trustee’s trustworthiness along specific dimensions. In IS, trust has traditionally been studied in interpersonal contexts, including organizations, and in relation to IT artefacts (Söllner et al., 2016; Thiebes et al., 2021). Depending on the nature of the trustee, different trusting beliefs become salient, including ability, benevolence, and integrity for individuals (Mayer et al., 1995) and functionality, helpfulness, and reliability for technology (McKnight et al., 2011; Söllner et al., 2013). For AI systems, additional antecedents of trust have been identified that vary depending on their form and level of intelligence (Glikson & Woolley, 2020; Lee & See, 2004), while established antecedents from both interpersonal and technological trust continue to apply (Thiebes et al., 2021).

Trust is dynamic and must be maintained once it has been established (Söllner & Pavlou, 2016). Building on this assumption, they developed a trust lifecycle model. After initial trust is formed, repeated interactions may stabilize it, but breaches can erode this trust, necessitating processes of repair before stability can be regained. In line with Kim et al. (2009), a trust violation occurs when an incident reduces trustor’s trusting beliefs about the trustee or the trustor’s trusting intentions (McKnight et al., 1998). Importantly, this reduction is not a direct consequence of the incident itself but results from how the trustor interprets and evaluates it. We conceptualize this process as sensemaking through which individuals interpret and give meaning to ambiguous or complex situations, enabling them to construct a coherent understanding that guides action (Weick, 1995). In this context, sensemaking involves forming causal attributions about the incident, which shape both emotional responses and subsequent trust judgments (Tomlinson & Mayer, 2009; Weiner, 1985).

For example, a trust violation may occur when the trustor concludes that the trustee lacks competence (“It gave me wrong information”), integrity (“It misled me”), or benevolence (“It did not act in my interest”). Building on this, research that trust repair depends on how the trustor attributes the incident, specifically along the dimensions of locus, controllability, and stability, which determine whether the incident is viewed as a temporary mistake or a signal of deeper untrustworthiness (Tomlinson & Mayer, 2009). Extending this work, Tomlinson et al. (2021) propose a cognitive process model in which trust repair unfolds through sequential stages of sensemaking, emotional response, and evaluation of the violator’s subsequent actions. These insights emphasize that trust violations are not defined solely by the event itself but by how the trustor interprets and evaluates it, making attribution and meaning-making central mechanisms in determining whether trust can be repaired.

2.2 Trust Violation and Repair in AI and Autonomous Systems

Trust violations in AI and autonomous systems differ from interpersonal settings due to several structural features of these technologies. First, users often struggle to interpret AI behavior because these systems operate in complex and partially random environments and rely on multilayered, opaque decision-making processes. As Glikson and Woolley (2020, p. 628) note, AI behavior is not

deterministic, and “the logic behind each decision made is often poorly understood,” making it difficult for users to predict why a violation occurred or whether it will recur. Second, trust repair in autonomous systems shows mixed and sometimes contradictory effects. Pak and Rovira (2024) argue that no single strategy reliably restores trust across contexts and emphasize that the effectiveness of repair depends on an interaction between properties of the repair strategy and the user’s ability, motivation, and opportunity to process it. Third, users appear to update their trust in AI differently than in human trustees. Research suggests that initial trust in automation can be unrealistically high because users ascribe superior analytical capability to automated aids (Parasuraman & Manzey, 2010), yet this trust can also decrease sharply after a single failure, reflecting heightened sensitivity to perceived violations (Madhavan & Wiegmann, 2007). These dynamics indicate that trust violations in AI systems emerge and are repaired within a distinct psychological and technological landscape that shapes how users interpret failures and respond to repair attempts.

Furthermore, the importance of trust calibration in automated systems is discussed in the literature. Users sometimes overestimate the analytical capability and authority of automated systems, which can lead to inflated trust before any violation occurs (Parasuraman & Manzey, 2010). For this reason, trust repair in autonomous systems is not aimed at maximizing trust but at restoring it to an appropriate, well-calibrated level. In this view, repair actions are understood as system-initiated responses to violations that momentarily reduce trust, with the goal of bringing trust back in line with the system’s actual capabilities. (Pak & Rovira, 2024)

3 Methodology

Given the interdisciplinary nature of our research question, we conducted a systematic literature review to capture and analyze contributions from IS, HCI, HRI, psychology, and organizational behavior, in line with recommendations for cross-disciplinary synthesis (Paré et al., 2015). Following established classifications (Cooper, 1988), we adopt a concept-centric, integrative review approach aimed at synthesizing and structuring existing conceptualizations of trust violations and their repair. We followed established guidelines for conducting SLRs (Page et al., 2021; vom Brocke et al., 2015; Webster & Watson, 2002) to ensure methodological rigor and transparency. A literature review was chosen because research on trust violations and trust repair is scattered across disciplines and has only recently begun to emerge within IS (Konopka & Wiesche, 2025), while adjacent fields, such as psychology, sociology, management, HCI, and HRI offer rich conceptual and empirical accounts of how trust violations are interpreted and addressed.

Since these insights from adjacent fields might be valuable for IS research, we performed an extensive search across five major bibliographic databases – ACM Digital Library, AISel, EBSCO, IEEE Xplore, and APA PsycInfo – and additionally reviewed journals included in the Senior Scholars’ List of Premier Journals, applying the search string to titles, abstracts, and keywords. To identify relevant studies, we constructed a search string incorporating multiple terms related to trust violations and repair. Drawing on terminology from recent publications, our final search string was:

“trust violation” OR “trust breach” OR “loss of trust” OR “broken trust” OR “trust erosion” OR “trust failure” OR “trust repair” OR “trust recovery” OR “trust restoration” OR “rebuilding trust” OR “regaining trust” OR “reestablishing trust”

To determine which studies to include in our review, we applied a set of predefined inclusion and exclusion criteria. We included papers that examined trust violations, loss of trust, trust erosion, or trust repair and that involved a sociotechnical configuration in which technology was directly involved in the trust relationship or violation incident. Specifically, this comprises cases where (1) the technology itself acted as the trustee, (2) the trust relationship was mediated through a technological system, or (3) the violation originated from a technological incident. Eligible studies offered either empirical findings or conceptual or theoretical insights related to trust breakdown or repair. We considered only peer-reviewed journal articles and full conference papers written in English. This approach served as a quality assurance mechanism by restricting the sample to peer-reviewed publications in established academic outlets. We excluded studies that focused solely on interpersonal or organizational trust without

technological involvement, that addressed trust without discussing violation or repair, or that conceptualized trust as a purely technical property (e.g., cryptographic or protocol trust) without user interaction or a sociotechnical perspective. We further excluded non-peer-reviewed work, literature reviews, and duplicate records.

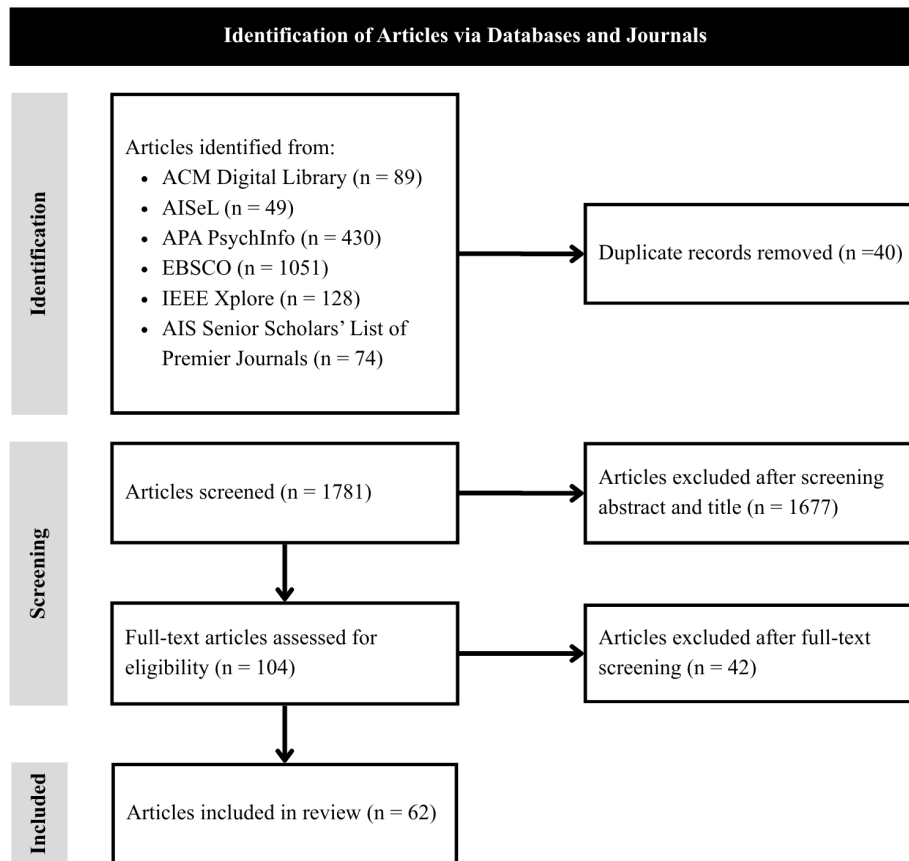


Figure 1. Literature Search Process

Our search across AISeL, APA PsycInfo, EBSCO, ACM Digital Library, IEEE Xplore, and the AIS Senior Scholars' List of Premier Journals yielded a total of 1,821 records, from which 40 duplicate records were removed. We first screened titles and abstracts of the remaining 1781 articles using the predefined inclusion and exclusion criteria, which reduced the set to 104 potentially relevant articles. Afterwards, we evaluated these articles based on their full text and excluded 42 papers based on our predefined inclusion and exclusion criteria. Our final sample included 62 articles (Figure 1).

To systematically analyze the final sample, we developed a concept matrix following the approach proposed by vom Brocke et al. (2015). We began by extracting all concepts related to trust violations and repair from each article and iteratively organizing them into thematic clusters. Through this iterative coding process, four overarching process stages emerged inductively from the literature: For each article, we coded all relevant details within these emerging categories, including the nature and context of the violation, the factors shaping its interpretation, the actions taken in response, and the resulting trust outcomes. As new concepts appeared, the matrix was continuously refined to ensure comprehensive coverage and conceptual consistency. This structured yet inductive approach enabled a transparent comparison across studies and provided the foundation for the process-oriented framework developed in the results.

4 Results

Our analysis of 62 articles shows that, despite fragmentation across domains, research on trust violations converges around a shared process. We synthesize this process into four stages: violation incident, trustor sensemaking, repair attempt, and recalibrated trust. The following subsections synthesize the concepts for each stage and identify key patterns, forming the empirical foundation for the Sociotechnical Trust Violation and Repair Process Framework.

4.1 Violation Incident

The violation incident captures what objectively happened before any sensemaking by the trustor. Across the reviewed studies, these events vary widely in form but converge around three underlying dimensions: (1) the type of incident that occurred, (2) the relationship between trustor and trustee, and (3) the context in which the incident took place. Across studies, violation incidents consistently fell into four recurring categories that we outline below.

Performance failures were the most common type of violation incidents and refer to situations in which a system produces incorrect outputs, behaves unreliably, or malfunctions (e.g. Hald et al., 2021; Han & Ko, 2025; Martell et al., 2025; Tan & Hao, 2023).

Privacy and security breaches involve the unauthorized exposure, misuse, or compromise of personal or sensitive information, predominantly in platform-based environments such as social networks (e.g. Mamonov & Koufari, 2014) and eCommerce (e.g. Muzatko & Bansal, 2024).

Expectation or contractual violations occur when systems fail to meet previously established goals, a pattern frequently observed in HRI (e.g. Li & Lee, 2025; Lyons et al., 2023) and organizational IS settings (e.g. Ngo-Ye et al., 2017).

Moral or ethical violations refer to incidents involving deception, norm violations, or behavior perceived as morally inappropriate (e.g. Phillips & Malle, 2025).

Most studies conceptualized trust violations as occurring in individual-technology relationships, but the identity of the trustee varied across domains. HRI was particularly dominant (e.g. Phillips & Malle, 2025). Work on automated or algorithmic systems (e.g. Dlodlo, 2025) similarly treated the system itself as the trustee. In contrast, eCommerce and financial services studies consistently positioned the organization behind the interface as the trustee, even when the violation was enacted by a technological artifact (e.g. Mamonov & Koufari, 2014; Muzatko & Bansal, 2024). Only a few papers examined multi-actor settings such as human-robot teams (e.g. Schelble et al., 2024).

Three empirical domains accounted for most violation incidents: (1) HRI, (2) automated and algorithmic systems, and, (3) platform-based environments such as eCommerce, financial services, and social networks. Beyond these, only scattered work examined other application areas such as cloud computing or autonomous driving. This concentration indicates that current knowledge about trust violations is shaped predominantly by robotics and platform technologies.

4.2 Sensemaking of Violation incident

The violation incident alone does not necessarily lead to a trust violation. In line with Kim et al. (2009), a trust violation occurs only when an incident reduces the trustor's trusting beliefs or their trusting intentions. This means that the trustor first interprets why the violation occurred and then evaluates whether their trusting beliefs are affected.

Attribution is a critical part of this sensemaking process. Building on attribution theory (Weiner, 1985), prior work suggests that causal attributions, specifically locus, controllability, and stability, shape whether a violation undermines perceived trustworthiness and signals the need for repair (Tomlinson & Mayer, 2009). For example, Jensen et al. (2019) examined whether users attribute blame for system errors to themselves, to system developers, or to external third-party algorithms. Research on e-commerce shows that trust violations can be attributed either to a lack of integrity or to a lack of competence, with buyers and sellers often disagreeing on which attribution is appropriate (Deng et al.,

2022). Trustworthiness judgments in online marketplaces similarly depend on whether users interpret a negative event as competence- or morality-based, with this attribution shaping how they evaluate the violator's response (Utz et al., 2009).

Contextual factors influencing sensemaking affect these attributions and the extent to which trustworthiness is perceived as violated. De Visser et al. (2016), found that anthropomorphism, which refers to the attribution of human traits to non-human entities such as technological systems (Fink, 2012), increases trust resilience in cognitive agents, and Dlodlo (2025) showed that anthropomorphic customer service agents can create a “trust shield” that buffers the perceived severity of service failures. Kahr et al. (2024) demonstrated that the timing of an error in AI advice affects how strongly trust and reliance decline, with early errors leading to more pronounced trust loss. Cultural background can also play a role: Zhang and Lee (2025) found that robotic failures had more negative effects on Korean users compared to users in the United States.

Across the reviewed studies, repair mechanisms differ not only in form but also in their effectiveness and the type of trust recovery they enable. Verbal strategies such as apologies, explanations, and excuses often lead to partial or context-dependent restoration, whereas more substantive responses, such as justifications or action-driven repairs, can contribute to stronger recovery of trustworthiness perceptions. At the same time, some strategies, such as re-ascription, involve trade-offs across trust dimensions rather than full restoration. These findings suggest that repair attempts do not uniformly restore trust but instead shape how trust is recalibrated, as examined in the following section.

4.3 Repair Attempt

The repair stage concerns whether and how the trustee responds to a violation, including whether action is taken at all (Li & Lee, 2025) and what triggers such responses, with prior research showing that repair may be *triggered* immediately (Luo et al., 2021) or only after the violation is voiced by the victim (Matzat & Snijders, 2012).

Apologies are among the most frequently studied trust repair mechanisms, yet their effectiveness is conditional rather than uniform. While they can contribute to restored trust (Bansal & Zahedi, 2015), their impact depends on factors such as attributional alignment: Cui et al. (2018) found that internal-attribution apologies are more effective for integrity-based violations, whereas external-attribution apologies are more effective for competence-based violations. Beyond attributional alignment, studies show that the effectiveness of apologies is shaped by contextual factors such as trustor's attitude (Esterwood & Robert, 2022). However, apologies often lead only to partial or dimension-specific restoration (Wan et al., 2023) and may fail to prevent reduced or even broken trust in safety-critical contexts (Liu et al., 2024).

Explanations are another frequently examined repair mechanism, but their effectiveness varies across contexts. In human-robot collaboration, Hald et al. (2021) found that neither explanations nor explanations combined with solutions increased trust after a robot mistake, despite users reporting that the information was helpful. In contrast, research in social commerce demonstrates a more robust effect: Wan et al. (2023) showed that explanation strategies most effectively reduced perceived service problems and helped restore benevolence, integrity, and competence beliefs following negative reviews. Together, these findings illustrate that the role of explanations in trust repair is context-dependent and not uniformly effective across domains.

Justifications represent a related but distinct repair mechanism compared to explanations. While explanations focus on why an action occurred, justifications also communicate the norms and values that guided the agent's decision. Phillips and Malle (2025) investigated this in the context of norm violations by social robots and found that justifications mitigated moral disapproval and restored perceived trustworthiness more effectively than explanations. Their experiments showed that people may still blame the robot for the specific norm violation but also appreciate the integrity signaled by the justified decision.

Re-ascription refers to repair attempts in which the violator disputes the original attribution to a specific trust dimension of the violation and instead re-attributes it. Deng et al. (2022a) show that this strategy

can contribute to restored trust in integrity when the violation is re-ascribed to competence or benevolence, but only for high-reputation. However, this strategy comes at a cost: acknowledging a competence- or benevolence-related deficiency reduces corresponding competence or benevolence beliefs regardless of reputation. Their findings suggest that re-ascription can be an effective but reputation-dependent repair strategy that involves systematic trade-offs across trust dimensions.

Excuses differ from both justifications and re-ascriptions. Unlike justifications, which defend the action by appealing to norms and values the agent aimed to uphold, excuses acknowledge the violation but attempt to reduce blame by denying intent or emphasizing temporary error. In this way, excuses share some similarity with re-ascriptions because both shift attention away from stable moral character, but excuses do so more weakly and without offering a clear causal alternative. Empirical evidence shows that excuses typically lead to only partial trust repair and are less effective in automated contexts (Langer et al. 2022).

Promises to improve future behavior represent another repair mechanism examined in the literature. In HRI interaction, Garcia Goo et al. (2025) show that trust lost after a navigational error could be effectively restored when a robot communicated its ability to learn from the mistake, especially when combined with an apology. In human-AI decision-making, Pareek et al. (2024) compare several trust repair strategies and found that promises were moderately effective for trust restoration, although less effective than model updates and apologies. Across these studies, communicating future improvement functions as a forward-looking signal that can help rebuild trust after a violation.

Action-driven forms of trust repair go beyond verbal repair strategies. In e-commerce, compensation represents a frequently tested mechanism. Wang et al. (2024) show that financial overcompensation was more effective than equal compensation in restoring trust. In human-robot collaboration, repair can take the form of providing the correct solution or correcting the error directly. Luo et al. (2021) demonstrate that prompt mistake correction, localized through human-to-robot attention transfer and executed immediately, effectively restored human trust across different types of robot errors. Together, these findings show that action-driven mechanisms can support trust repair, particularly when they provide concrete correction or compensation rather than solely verbal accounts.

Across the reviewed studies, repair mechanisms differ not only in form but also in their effectiveness and the type of trust recovery they enable. Verbal strategies such as apologies, explanations, and excuses often lead to partial or context-dependent restoration (e.g. Wan et al., 2023), whereas responses, such as justifications or action-driven repairs, can contribute to stronger recovery of trustworthiness perceptions (e.g. Luo et al., 2021; Phillips & Malle, 2025). At the same time, some strategies, such as re-ascription, involve trade-offs across trust dimensions rather than full restoration (Deng et al., 2022). These findings suggest that repair attempts do not uniformly restore trust but instead shape how trust is recalibrated, as examined in the following section.

4.4 Recalibrated Trust

Trust recalibration captures how trust changes following repair attempts, ranging from breakdown to full or even enhanced recovery. As shown in the preceding section, repair mechanisms vary considerably in their effectiveness, and these differences are reflected in the diverse patterns of trust recalibration observed across studies.

Broken trust was documented by several papers, where trust could not be recovered despite repair attempts. Schelble et al. (2024) show that neither apologies nor denials restored trust after ethical violations by an autonomous teammate. *Reduced trust*, similarly was reported in some studies. Martell et al. (2025) similarly report that trust in an LLM remained durably reduced after an early error, with mitigation strategies failing to produce recovery. While both outcomes reflect negative trust recalibration, they differ in degree and recoverability: reduced trust describes a decline in trust that remains adjustable and may partially recover over time, whereas broken trust refers to cases in which trust collapses and cannot be restored despite repair attempts.

Restored trust, in contrast, where trust returned to pre-violation levels following an effective repair attempt, was demonstrated in other studies. For example, Luo et al. (2021) show that prompt mistake correction in human-robot collaboration can fully restore trust after performance errors.

Enhanced trust, where trust exceeded its initial baseline following a repair attempt, was reported in some studies. Han and Ko (2025), in a financial advisory context, found that post-error explanations not only facilitated recovery but produced trust levels higher than those observed before the error. In addition to overall trust, recalibration can occur at the level of specific trustworthiness dimensions, and several studies show that recovery may be context-dependent, with trust restored for the immediate task but not necessarily generalized to broader settings.

Specific trustworthiness dimensions matter for trust repair and the reviewed studies show that these dimensions do not recover uniformly. Alarcon et al. (2024) demonstrate that violations affect performance, process, and purpose perceptions differently: while performance perceptions recovered after repair, process and purpose perceptions declined following a distrust behavior and did not return to baseline, suggesting that deeper relational and moral evaluations are more resistant to repair. Similar patterns emerge in service and platform contexts. Wan et al. (2023) show that negative reviews can reduce benevolence, integrity, and competence beliefs, and that repair strategies restore these dimensions to varying degrees. Apologies primarily improve attitude-related perceptions, explanations repair all three dimensions most effectively, and compensation restores perceptions related to service quality. These findings highlight that trustworthiness recalibration is dimension-specific and that the effectiveness of repair strategies depends on the particular belief that was violated.

Context matters for trust recalibration as several studies indicate. Textor et al. (2022) show that after ethical violations by an autonomous teammate, participants' trust in the agent's moral integrity declined, yet they continued to trust the agent to perform task-related actions. This suggests that trust can fragment across domains, with ethical trust remaining damaged while functional trust remains usable. Similar patterns emerge in performance-focused contexts. Luo et al. (2021) demonstrate that prompt mistake correction restores trust for the specific collaborative task, but trust remains fragile and may not generalize to new situations. These findings highlight that trust repair may succeed in restoring trust for particular tasks or contexts while broader or more abstract trustworthiness evaluations remain impaired.

5 Sociotechnical Trust Violation and Repair Process

Based on the insights from our review, we propose a Sociotechnical Trust Violation and Repair Process Framework for analyzing trust violations and repair attempts in sociotechnical systems, both in existing research and in empirical cases. Although prior work has examined individual aspects of trust violations, these insights remain scattered across domains and are often not connected to a shared conceptual structure. Our synthesis shows that trust violations unfold through a recurring sequence consisting of four stages: the violation incident, the trustor sensemaking of the event, the violator reaction, and the resulting trust outcome. To support systematic analysis across different contexts, we complement this framework with a set of guiding questions that specify what needs to be understood at each stage. Together, the framework and the questions offer a coherent analytical lens that can be applied to empirical cases of trust violation, regardless of the specific technology, domain, or actor involved.

Figure 2 illustrates the core logic of the Sociotechnical Trust Violation and Repair Process. Rectangular elements represent actions by the trustor or violator, whereas circular elements denote events, states, or outcomes within the process. A trust violation originates in an observable violation incident. Whether this incident becomes a perceived trust violation depends on how the trustor makes sense of it. If the event is evaluated as violating one or more trusting beliefs, resulting in a perceived trust violation, it may trigger a repair attempt initiated by the trustor. Independently of this evaluation, the system may also initiate a repair attempt in response to the violation incident itself. These repair attempts may succeed, fail, or not occur at all, leading to a state of recalibrated trust that may reflect restored, reduced, enhanced, or broken trust. If the trustor does not interpret the incident as violating trusting beliefs, the process ends with no trust violation.

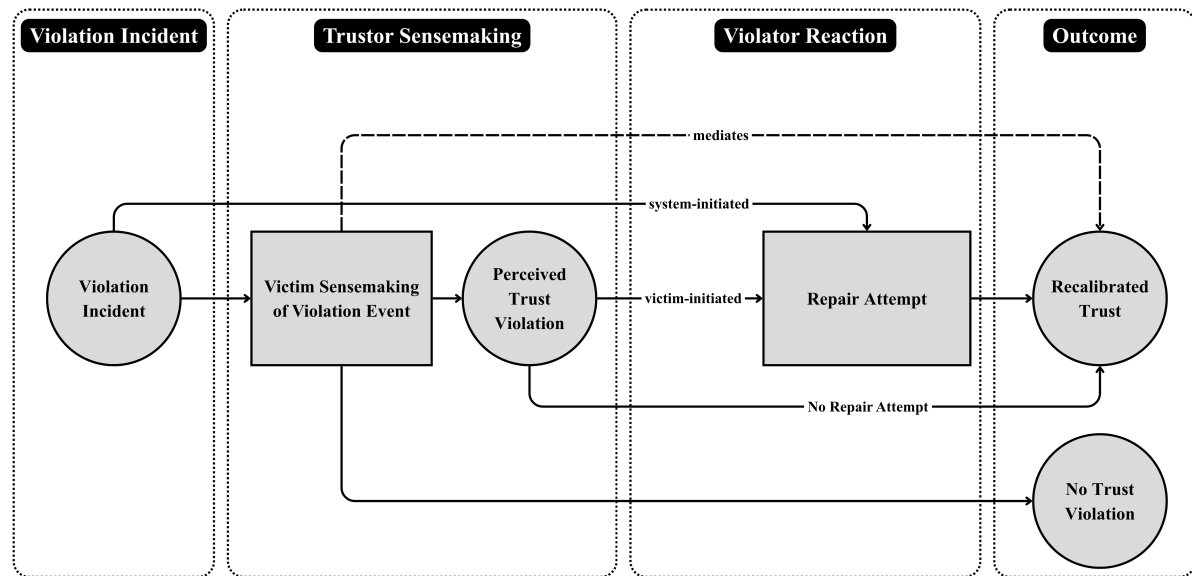


Figure 2. Sociotechnical Trust Violation and Repair Process Framework

While the Sociotechnical Trust Violation and Repair Process Framework structures how trust violations unfold, the guiding questions in Table 1 specify what an analyst must pay attention to understand a trust violation in a sociotechnical system. The questions are derived from the concepts presented in Section 4 and are organized according to the four stages of the Sociotechnical Trust Violation and Repair Process Framework. They are not intended to provide a structured way to analyze individual cases and compare studies that use different terminology.

Table 1. Guiding Questions for Analyzing Trust Violations and Repair in Sociotechnical Systems

Violation Incident	Victim Sensemaking	Violator Reaction	Outcome
What happened?	How does the violation incident affect the trustor's trusting beliefs?	Does the violator take any action to repair trust?	Is the repair attempt effective in repairing trust?
Between whom?	What factors mediate how the trustor interprets the violation incident?	What triggers the repair attempt?	To what extent has the trust level changed compared to before the violation?
In what context did the violation incident occur?	How does the trustor attribute blame?	What action does the violator take?	To what extent have the trusting beliefs of the trustor changed?
			Has the trust outcome only changed in relation to a specific context?

At the level of the “Violation Incident,” the questions clarify what objectively happened, between which actors, and in what context, ensuring that the starting point of a trust violation is described precisely. For “Trustor Sensemaking,” they focus on how the event is interpreted, which trust beliefs are affected, and which factors shape these interpretations, such as attributions or contextual cues. For “Violator Reaction,” they identify whether a repair attempt occurred, what triggered it, and how the violator responded. In the “Outcome” stage, they address whether trust was repaired, how it changed relative to the pre-violation state, which dimensions shifted, and whether recovery was global or context-specific.

Together, these questions make the framework actionable by guiding analysis across the full sequence of a trust violation and enabling structured comparison across cases.

6 Discussion

This study set out to answer the research question of how trust violations and their repair have been conceptualized across sociotechnical systems, with particular attention to the relationship between trustor sensemaking and violator reactions. Synthesizing insights from 62 studies, we developed a Sociotechnical Trust Violation and Repair Process framework. Building on this framework, we derive theoretical contributions, practical implications, and directions for future research.

6.1 Theoretical Implications

First, our findings highlight the central role of trustor sensemaking in trust violation and repair processes. Our synthesis shows that violation incidents do not directly translate into trust breakdown. Instead, whether an incident becomes a trust violation depends on how the trustor interprets it through attribution and evaluation processes. By consolidating insights across disciplines, we position sensemaking as the key mechanism linking violation incidents to trust outcomes, thereby shifting attention from system behavior alone to the interaction between system behavior and user cognition.

Second, this study provides a stage-based analytical lens for examining trust violations in sociotechnical systems. By organizing the process into four analytically distinct but interrelated stages - violation incident, trustor sensemaking, repair attempt, and recalibrated trust - the framework enables researchers to disentangle how trust violations unfold and where specific mechanisms operate. The accompanying guiding questions further operationalize these stages by specifying what needs to be examined at each point in the process, such as how violations are interpreted, how repair is enacted, and how trust changes as a result.

Third, our findings suggest a reframing of trust repair outcomes as processes of calibration rather than simple restoration. Our synthesis shows that post-violation trust can take multiple forms, including reduced, restored, enhanced, or broken trust, and that these outcomes may differ across trust dimensions and contexts. This highlights that trust repair does not necessarily return trust its pre-violation level, but instead results in a recalibrated state. Importantly, this perspective also suggests that in some cases, a violation may signal that prior trust was inflated, and a subsequent reduction in trust may lead to a more appropriate alignment between user expectations and system capabilities. In this sense, trust violations and repair can be understood as mechanisms through which users adjust their reliance on sociotechnical systems. This interpretation aligns with work on human-automation interaction, which emphasizes the importance of calibrated trust for appropriate reliance (Lee & See, 2004).

6.2 Practical Implications

A central implication of our findings is that trust management in AI-enabled systems must place the user at the center. Specifically, practitioners need to understand how users interpret system behavior and attribute failures, as these processes shape whether an incident is perceived as a trust violation and how subsequent repair efforts are evaluated. This is particularly important for selecting appropriate repair mechanisms, as their effectiveness depends on how the violation is interpreted; for example, apologies tend to be more effective for integrity-based violations (Cui et al., 2018). Taken together, this suggests that trust formation and failure handling should not be treated as isolated events, but as interconnected processes shaped by user perceptions and responses.

In the same vein, trust violations and repair should be analyzed across the entire process rather than as isolated moments. To support this, the guiding questions developed in this study provide a structured way to analyze trust violations from a process perspective, helping practitioners to identify how incidents are interpreted and whether repair efforts succeed or fail. By making the trustor's perspective more explicit, this approach may also help practitioners anticipate potential trust breakdowns and identify opportunities for preventing them.

6.3 Directions for Further Research

Building on the four stages of our framework, we highlight several directions (D) for future research.

Violation Incident: In IS, empirical work on trust violations has been concentrated primarily in eCommerce contexts (Deng et al., 2022; Muzatko & Bansal, 2024), where violations typically involve organizations. In contrast, much less is known about violations that arise in interactions between users and technological artifacts, even though AI-enabled and autonomous systems increasingly operate as semi-independent actors within organizational and consumer settings (Glikson & Woolley, 2020). As these systems take on more agentic and decision-making roles, IS research must develop a clearer understanding of what kinds of incidents users construe as trust violations. Therefore, we propose the following direction:

D1: Examine how trust violation incidents emerge in human-technology interactions across different IS contexts, beyond traditional eCommerce settings.

Trustor Sensemaking: Whether a violation has occurred hinges on how the user interprets the event – the victim sensemaking. This aligns with theoretical work in trust violation and repair, which emphasizes sensemaking, causal attributions, and interpretations of intent or negligence as the mechanisms through which violations become meaningful (Tomlinson et al., 2021; Tomlinson & Mayer, 2009). Yet in IS research, trust is often approached from a design-oriented perspective aimed at engineering system features that elicit desired levels of trust (e.g., Qiu & Benbasat, 2005). Shifting attention toward the trustor foregrounds how users' expectations, needs, dispositions, and interpretive processes shape when they perceive an incident as a violation, how severe they judge it to be, and whether they believe repair is necessary. Therefore, we propose the following directions:

D2.1: Investigate how trustors interpret and attribute causes to system failures, and under which conditions incidents are construed as trust violations.

D2.2: Examine how contextual cues and individual differences shape the trustor's sensemaking process and the perceived need for repair.

Repair Attempt: Our review highlights that many repair strategies studied in AI and robotics – such as apologies, justifications, or excuses – are inherently anthropomorphic. Pak and Rovira (2024) explicitly argue that these are forms of social signaling that leverage human social expectations. Yet, the mixed evidence on their effectiveness suggests that anthropomorphic repair cues are not universally beneficial and may interact with prior expectations, system transparency, and user motivation. At the same time, AI systems have the ability to provide solution-based repair; for example, correcting an error immediately or regenerating the correct answer. Our review indicates that this strategy, while rare in among our reviewed studies, may be particularly salient for AI. Since errors are often stochastic, systems can rapidly recompute correct outcomes, correction can serve as a concrete demonstration of reliability. Future IS research should investigate how solution-based, self-correcting, or self-updating repair strategies influence trust differently from human-like apologies or explanations. Therefore, we propose the following directions:

D3.1 Study which anthropomorphic cues (e.g., apologies, promises) are interpreted as credible forms of repair.

D3.2 Investigate how different forms of anthropomorphism shape trustor attributions after a violation.

D3.3: Study the effectiveness of action-based repair strategies (e.g., prompt correction) relative to verbal strategies.

Recalibrated Trust: Current IS research often treats trust repair as a success when trust increases. However, in AI contexts, especially with high-performing or opaque systems, trust must not simply be restored but appropriately calibrated. Prior work in autonomy research show that users frequently overestimate AI systems' capabilities prior to an violation (Parasuraman & Manzey, 2010). At the same time, trust may decrease sharply after a single failure (Madhavan & Wiegmann, 2007). This asymmetry suggests that the optimal post-repair outcome is not necessarily the original level of trust, but an alignment between user beliefs and system capabilities (Pak & Rovira, 2024). In this sense, both trust violations and repair strategies may serve as opportunities for recalibration.

As shown by Textor et al. (2022), users may lose trust in an AI system's ethical judgment yet continue to rely on its task performance. This suggests that trust may not recover or deteriorate uniformly across domains of use. For general-purpose AI systems that operate across multiple contexts, this fragmentation becomes especially important and warrants dedicated research attention.

D4.1: Explore how violation incidents and repair strategies can be designed to support appropriate trust calibration.

D4.2: Investigate how trust recalibration unfolds across contexts, particularly for general-purpose AI systems that operate across multiple domains of user expectations.

7 Conclusion and Limitations

This paper synthesizes research on sociotechnical trust violations, integrating insights from 62 studies across domains into a unified, process-oriented framework. It shows that trust violations follow a recurring sequence – violation incident, victim sensemaking, violator reaction, and trust recalibration – consistent across disciplines despite differing vocabularies. By placing the trustor's sensemaking at the center, the framework advances IS trust research beyond design-centric perspectives and provides a foundation for future theorizing on human-AI trust dynamics. Practically, the framework and accompanying guiding questions offer a structured tool for diagnosing trust breakdowns in real-world systems, informing the design of repair strategies, and supporting audits and governance processes aimed at maintaining calibrated trust in AI and automated systems.

Despite following systematic procedures in identifying and selecting the literature, our review has several limitations. First, our search strategy was intentionally broad, aiming to capture conceptual diversity rather than focus on a single application domain. As a result, relevant work that adopts different terminology or appears in adjacent fields may not have been included. Second, while the review process was systematic, we did not assess the methodological quality of individual studies, which limits our ability to weigh empirical findings relative to their rigor. Third, our synthesis relies on the definitions and categorizations present in the included literature; using alternative conceptualizations of trust or trust violations might yield different emphases or groupings.

References

- Alarcon, G. M., Lyons, J. B., Hamdan, I. aldin, & Jessup, S. A. (2024). Affective responses to trust violations in a human-autonomy teaming Context: Humans versus robots. *International Journal of Social Robotics*, 16(1), 23–35.
- Bachura, E., Valecha, R., Chen, R., & Rao, H. R. (2022). The OPM Data Breach: An Investigation of Shared Emotional Reactions on Twitter. *Management Information Systems Quarterly*, 46(2), 881–910. <https://doi.org/10.25300/MISQ/2022/15596>
- Bansal, G., & Zahedi, F. M. (2015). Trust violation and repair: The information privacy perspective. *Decision Support Systems*, 71, 62–77. <https://doi.org/10.1016/j.dss.2015.01.009>
- Chinman, L. (2025, August 15). *Influencer Breaks Down in Tears After Trusting ChatGPT for Travel Advice, and Then Missing Flight*. People.Com. <https://people.com/influencer-misses-flight-after-trusting-chatgpt-visa-advice-11791860>
- Choi, B. C. F., Kim, S. S., & Jiang, Z. (Jack). (2016). Influence of Firm's Recovery Endeavors upon Privacy Breach on Online Customer Behavior. *Journal of Management Information Systems*, 33(3), 904–933. <https://doi.org/10.1080/07421222.2015.1138375>
- Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society*, 1(1), 104. <https://doi.org/10.1007/BF03177550>
- Cui, Y., Zhang, X., Peng, X., & Chu, J. (2018). How to use apology and compensation to repair competence-versus integrity-based trust violations in e-commerce. *Electronic Commerce*

- Research & Applications*, 32, 37–48. (133392687).
<https://doi.org/10.1016/j.elerap.2018.11.001>
- De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., 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. <https://doi.org/10.1037/xap0000092>
- Deng, H., Wang, W., & Lim, K. (2022). Repairing Integrity-Based Trust Violations in Ascription Disputes for Potential E-Commerce Customers. *Management Information Systems Quarterly*, 46(4), 1983–2014.
- Dlodlo, N. (2025). Anthropomorphic service recovery: The panacea following service failure of automated CSAs. *Acta Commercii*, 25(2), 1–10. <https://doi.org/10.4102/ac.v25i2.1348>
- Emery, F. (2016). Characteristics of Socio-Technical Systems. In *The Social Engagement of Social Science, a Tavistock Anthology, Volume 2* (pp. 157–186). University of Pennsylvania Press. https://www.degruyterbrill.com/document/doi/10.9783/9781512819052-009/html?srsId=AfmBOoqw_4EI9cte1bVXsa3PweP443jAh6nuJtHPMUMRAo4KbB1CnHqu
- Esterwood, C., & Robert, L. P. (2022). Having the Right Attitude: How Attitude Impacts Trust Repair in Human-Robot Interaction. *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction, HRI '22*, 332–341.
- Esterwood, C., & Robert, L. P. (2023). Three Strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness. *Computers in Human Behavior*, 142, 107658. <https://doi.org/10.1016/j.chb.2023.107658>
- Fink, J. (2012). Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction. In S. S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, & M.-A. Williams (Eds.), *Social Robotics* (Vol. 7621, pp. 199–208). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34103-8_20
- Garcia Goo, H., Schadenberg, B. R., Kolkmeier, J., Truong, K. P., & Evers, V. (2025). Being Sorry is the Hardest Thing: How Robots can Apologize and Learn from Mistakes to Restore People’s Trust. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '25*, 1–9. <https://doi.org/10.1145/3706599.3719739>
- Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, 14(2), 627–660. <https://doi.org/10.5465/annals.2018.0057>
- Hald, K., Weitz, K., André, E., & Rehm, M. (2021). “An Error Occurred!”—Trust Repair With Virtual Robot Using Levels of Mistake Explanation. *Proceedings of the 9th International Conference on Human-Agent Interaction, HAI '21*, 218–226. <https://doi.org/10.1145/3472307.3484170>
- Han, J., & Ko, D. (2025). Trust Formation, Error Impact, and Repair in Human–AI Financial Advisory: A Dynamic Behavioral Analysis. *Behavioral Sciences*, 15(10), 1370. <https://doi.org/10.3390/bs15101370>
- Jensen, T., Albayram, Y., Khan, M. M. H., Fahim, M. A. A., Buck, R., & Coman, E. (2019). The Apple Does Fall Far from the Tree: User Separation of a System from its Developers in Human-Automation Trust Repair. *Proceedings of the 2019 on Designing Interactive Systems Conference, DIS '19*, 1071–1082. <https://doi.org/10.1145/3322276.3322349>
- Kahr, P. K., Rooks, G., Snijders, C., & Willemsen, M. C. (2024). The Trust Recovery Journey. The Effect of Timing of Errors on the Willingness to Follow AI Advice. *Proceedings of the 29th*

- International Conference on Intelligent User Interfaces, IUI '24*, 609–622. <https://doi.org/10.1145/3640543.3645167>
- 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. <https://doi.org/10.5465/amr.2009.40631887>
- Konopka, B., & Wiesche, M. (2025, December 14). Can “AI” Repair Trust? Comparing Human-Like and System-Like Can “AI” Repair Trust? Comparing Human-Like and System-Like Repair Strategies. *ICIS 2025 Proceedings*. International Conference on Information Systems (ICIS).
- Langer, M., König, C. J., Back, C., & Hemsing, V. (2023). Trust in Artificial Intelligence: Comparing Trust Processes Between Human and Automated Trustees in Light of Unfair Bias. *Journal of Business and Psychology*, 38(3), 493–508. <https://doi.org/10.1007/s10869-022-09829-9>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- Li, M., & Lee, J. D. (2025). Purpose Outweighs Performance: Trust Drops More When AI Teammates Fail to Cooperate, but Explanations Can Repair It. *International Journal of Human-Computer Interaction*, 1–22. <https://doi.org/10.1080/10447318.2025.2539449>
- Liu, P., Chu, Y., Wang, G., & Xu, Z. (2024). To Err is Automation: Can Trust be Repaired by the Automated Driving System After its Failure? *IEEE Transactions on Human-Machine Systems*, 54(5), 508–519. <https://doi.org/10.1109/THMS.2024.3434680>
- Luo, R., Huang, C., Peng, Y., Song, B., & Liu, R. (2021). Repairing Human Trust by Promptly Correcting Robot Mistakes with An Attention Transfer Model. *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, 1928–1933. <https://doi.org/10.1109/CASE49439.2021.9551502>
- Lyons, J. B., Hamdan, I. aldin, & Vo, T. Q. (2023). Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior*, 138, 1–11.
- Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, 8(4), 277–301. <https://doi.org/10.1080/14639220500337708>
- Maehigashi, A., Kubo, K., Nungduk, Y., & Yamada, S. (2025). Effects of Robot Bowing during Apology on Trust Repair. *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction, HRI '25*, 1478–1482.
- Mamonov, S., & Koufaris, M. (2014). The Impact of Perceived Privacy Breach on Sustainability of Social Networking Sites. *2014 47th Hawaii International Conference on System Sciences*, 1774–1784. <https://doi.org/10.1109/HICSS.2014.225>
- Martell, M. J., Baweja, J. A., & Dreslin, B. D. (2025). Mitigative Strategies for Recovering From Large Language Model Trust Violations. *Journal of Cognitive Engineering and Decision Making*, 19(1), 76–95. <https://doi.org/10.1177/15553434241303577>
- Matzat, U., & Snijders, C. (2012). Rebuilding Trust in Online Shops on Consumer Review Sites: Sellers’ Responses to User-Generated Complaints. *Journal of Computer-Mediated Communication*, 18(1), 62–79. <https://doi.org/10.1111/j.1083-6101.2012.01594.x>
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3), 709–734. <https://doi.org/10.2307/258792>
- McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, 2(2), 1–25. <https://doi.org/10.1145/1985347.1985353>

- McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). INITIAL TRUST FORMATION IN NEW ORGANIZATIONAL RELATIONSHIPS. *Academy of Management Review*, 23(3), 473–490.
- Muzatko, S., & Bansal, G. (2024). It pays to be forthcoming: Timing of data breach announcement, trust violation, and trust restoration. *Internet Research*, 34(5), 1629–1663. (179977950). <https://doi.org/10.1108/INTR-12-2021-0939>
- Ngo-Ye, T. L., NAZARETH, D. L., & CHOI, J. J. (2017). Trust in Cloud Computing: Maintaining Long-Term Relationships. *Journal of Information Technology Management*, 28(3), 25–38.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Pak, R., & Rovira, E. (2024). A theoretical model to explain mixed effects of trust repair strategies in autonomous systems. *Theoretical Issues in Ergonomics Science*, 25(4), 453–473. (178419271). <https://doi.org/10.1080/1463922X.2023.2250424>
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52(3), 381–410. <https://doi.org/10.1177/0018720810376055>
- Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, 52(2), 183–199. <https://doi.org/10.1016/j.im.2014.08.008>
- Pareek, S., Velloso, E., & Goncalves, J. (2024). Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, 546–561. <https://doi.org/10.1145/3630106.3658924>
- Phillips, E. K., & Malle, B. F. (2025). The power of justifications to repair human-robot trust, even under moral disagreement. *Scientific Reports*, 15(1), 34706. <https://doi.org/10.1038/s41598-025-17983-2>
- Qiu, L., & Benbasat, I. (2005). Online Consumer Trust and Live Help Interfaces: The Effects of Text-to-Speech Voice and Three-Dimensional Avatars. *International Journal of Human-Computer Interaction*, 19(1), 75–94. https://doi.org/10.1207/s15327590ijhc1901_6
- Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., & Freeman, G. (2024). Towards Ethical AI: Empirically Investigating Dimensions of AI Ethics, Trust Repair, and Performance in Human-AI Teaming. *Human Factors*, 66(4), 1037–1055. (35938319). <https://doi.org/10.1177/00187208221116952>
- Sebo, S. S., Krishnamurthi, P., & Scassellati, B. (2020). “I don’t believe you”: Investigating the effects of robot trust violation and repair. *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction, HRI '19*, 57–65.
- Söllner, M., Benbasat, I., Gefen, D., Leimeister, J. M., & Pavlou, P. A. (2018). Trust. In *MIS Quarterly Research Curations, Ashley Bush and Arun Rai, Eds., Http://Misq.Org/Research-Curations, Released: October 2016, Last Update: June 2018*.
- Söllner, M., Hoffmann, A., & Leimeister, J. M. (2016). Why different trust relationships matter for information systems users. *European Journal of Information Systems*. (world). <https://doi.org/10.1057/ejis.2015.17>

- Söllner, M., & Pavlou, P. A. (2016). *A Longitudinal Perspective on Trust in IT Artefacts*. *Research Papers*. ECIS 2016 Proceedings. https://aisel.aisnet.org/ecis2016_rp/52
- Söllner, M., Pavlou, P. A., & Leimeister, J. M. (2013). *Understanding Trust in IT Artifacts – A New Conceptual Approach* (SSRN Scholarly Paper No. 2475382). <https://doi.org/10.2139/ssrn.2475382>
- Tan, H., & Hao, Y. (2023). How does people’s trust in automated vehicles change after automation errors occur? An empirical study on dynamic trust in automated driving. *Human Factors & Ergonomics in Manufacturing & Service Industries*, 33(6), 449–463. <https://doi.org/10.1002/hfm.21001>
- Textor, C., Zhang, R., Lopez, J., Schelble, B. G., McNeese, N. J., Freeman, G., Pak, R., Tossell, C., & De Visser, E. J. (2022). Exploring the Relationship Between Ethics and Trust in Human–Artificial Intelligence Teaming: A Mixed Methods Approach. *Journal of Cognitive Engineering and Decision Making*, 16(4), 252–281. <https://doi.org/10.1177/15553434221113964>
- Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, 31(2), 447–464. <https://doi.org/10.1007/s12525-020-00441-4>
- Tomlinson, E. C., & Mayer, R. C. (2009). The Role Of Causal Attribution Dimensions In Trust Repair. *Academy of Management Review*, 34(1), 85–104. <https://doi.org/10.5465/amr.2009.35713291>
- Tomlinson, E. C., Nelson, C. A., & Langlinais, L. A. (2021). A cognitive process model of trust repair. *International Journal of Conflict Management (Emerald)*, 32(2), 340–360. <https://doi.org/10.1108/IJCMA-03-2020-0048>
- 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. <https://doi.org/10.2753/JEC1086-4415130304>
- vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. *Communications of the Association for Information Systems*, 37(1). <https://doi.org/10.17705/1CAIS.03709>
- Wan, Y., Zhang, Y., Wang, F., & Yuan, Y. (2023). Retailer response to negative online consumer reviews: How can damaged trust be effectively repaired? *Information Technology & Management*, 24(1), 37–53. (162181483). <https://doi.org/10.1007/s10799-022-00367-7>
- Wang, Q., Han, P., Hao, S., & Yang, G. (2024). Effect of financial overcompensation on consumer trust after e-commerce service failures: Evidence from event-related potentials. *Current Psychology*, 43(17), 15621–15632. (176782012). <https://doi.org/10.1007/s12144-023-05500-5>
- Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, 26(2), xiii–xxiii.
- Weick, K. E. (1995). *Sensemaking in Organizations* (1. [Dr.]). Sage.
- Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, 92(4), 548–573. <https://doi.org/10.1037/0033-295X.92.4.548>
- Zahedi, F. “Mariam,” & Song, J. (2008). Dynamics of Trust Revision: Using Health Infomediaries. *Journal of Management Information Systems*, 24(4), 225–248. <https://doi.org/10.2753/MIS0742-1222240409>
- Zhang, X., & Lee, S. K. (2025). “But You Shouldn’t Blame Me”: A cross-national comparison of the effects of performance failures and trust repairs in human-robot interactions. *J. Hum.-Robot Interact*. <https://doi.org/10.1145/3746461>