Manavest
Journal of Human Capital and Organizational Behavior

Vol. 1 l Issue 1 l April 2026
Pages 60-77

Trust Breakdown in AI-Driven Organizations: A Mechanism-Based Explanation

Usep Deden Suherma [1]*

***Corresponding Mail:**
sepds@uinsgd.ac.id

Article History:
Submitted: 19-12-2025
Approved: 23-02-2026
Published: 06-04-2026

OPEN ACCESS
Available at the open access journal:
https://sciedex.com/manavest

Abstrak

Artificial intelligence (AI) is increasingly embedded in organizational decision-making, promising enhanced efficiency, objectivity, and performance. However, growing evidence indicates a paradox in which the advancement of AI systems is accompanied by declining trust among organizational members. Existing research on trust, artificial intelligence, and organizational behavior remains fragmented, limiting its ability to explain how AI-driven systems reshape trust dynamics in complex work environments. This study addresses this gap by developing a mechanism-based conceptual framework that explains how structural characteristics of AI systems, including opacity, autonomy, and unpredictability, trigger psychological processes that lead to trust breakdown. Drawing on interdisciplinary insights, the framework identifies key mediating mechanisms such as cognitive uncertainty, perceived loss of control, identity threat, and fairness ambiguity, which translate system features into behavioral outcomes including distrust, resistance, and disengagement. The study further highlights the moderating role of organizational conditions, including transparency, leadership support, and human–AI interaction quality. By integrating previously disconnected research streams, this article advances a process-oriented understanding of trust in AI-driven organizations and provides a foundation for future empirical research and practical interventions in managing trust under digital transformation.

Keywords

artificial intelligence; organizational trust; trust breakdown; human–AI interaction; digital transformation

[1] Department of Sharia Financial Management, Faculty of Islamic Economics and Business, Sunan Gunung Djati State Islamic University, Bandung, Indonesia

1. Introduction

Artificial intelligence has become a central driver of contemporary organizational transformation, fundamentally reshaping how decisions are made, processes are optimized, and value is created. Organizations increasingly rely on AI-enabled systems to enhance efficiency, improve accuracy, and reduce human bias in decision-making processes. These systems are often portrayed as objective and superior alternatives to human judgment, capable of processing vast amounts of data and generating insights at unprecedented speed. The broader trajectory of digital transformation further amplifies this shift, embedding AI into core organizational functions and redefining work structures and managerial practices (Vial, 2019; Benbya *et al.*, 2021). As AI capabilities continue to advance, organizations are expected to become more rational, data-driven, and performance-oriented, reinforcing the belief that technological sophistication naturally leads to better organizational outcomes.

However, emerging evidence suggests a paradox that challenges this dominant narrative. Despite the increasing sophistication of AI systems, trust within organizations does not necessarily improve and may, in fact, deteriorate. The assumption that advanced technologies inherently foster confidence among organizational members is increasingly contested. Instead, AI systems often introduce opacity, reduce interpretability, and alter the locus of control in decision-making processes, creating conditions that complicate trust formation (Afroogh *et al.*, 2024; Benda *et al.*, 2022). As decision authority shifts from human actors to algorithmic systems, individuals are required to rely on outputs they may not fully understand, evaluate, or challenge. This shift generates uncertainty and raises fundamental questions about accountability, fairness, and legitimacy, thereby destabilizing traditional foundations of trust in organizational contexts.

This tension becomes more evident when examining empirical developments in digitally transformed organizations. Increasingly, employees exhibit distrust toward AI-based decision systems, particularly in high-stakes or ambiguous contexts. Rather than fully embracing algorithmic recommendations, individuals often question their validity, resist their implementation, or selectively disregard their outputs. Such reactions are not isolated incidents but reflect a broader pattern of skepticism toward algorithmic authority. Recent studies document how employees express concerns about the reliability and transparency of AI systems, leading to hesitation and reduced willingness to rely on algorithmic decisions (Cho *et al.*, 2025). Similarly, evidence from knowledge work environments indicates that interactions with AI systems can generate uncertainty and discomfort, especially when system outputs contradict human expectations or lack clear explanations (Brown *et al.*, 2025).

Moreover, these dynamics are increasingly linked to behavioral consequences that extend beyond immediate decision contexts. Employees who perceive AI systems as unreliable or opaque may experience diminished trust not only in the technology itself but also in the broader organizational structures that implement such systems. This erosion of trust can manifest in reduced engagement, increased resistance to technological change, and a decline in innovative behavior (Liu *et al.*, 2026). The persistence of these patterns suggests that distrust in AI-driven environments is not merely a transitional issue associated with early adoption but represents a deeper and more systemic organizational challenge. As AI becomes more embedded in everyday work practices, the implications of trust breakdown become more pronounced and consequential for organizational performance and sustainability.

Despite the growing recognition of these issues, the theoretical foundations for understanding trust in AI-driven organizations remain fragmented. Research on artificial intelligence has predominantly focused on technological capabilities, including algorithmic performance, accuracy, and efficiency. While these studies provide valuable insights into

system design and functionality, they often overlook the human and organizational dimensions of AI adoption. In contrast, the organizational behavior literature has extensively examined human attitudes, behaviors, and psychological processes in workplace settings, yet it rarely accounts for the unique characteristics of AI systems as non-human actors. Meanwhile, trust research has traditionally emphasized interpersonal relationships, conceptualizing trust as a relational construct grounded in perceptions of ability, benevolence, and integrity (Rousseau *et al*., 1998; Dirks & Ferrin, 2001).

Although recent studies have begun to explore trust in technological and algorithmic contexts, these efforts remain limited in their ability to integrate the complexities of AI-driven environments. For instance, research on algorithmic transparency and explainability highlights the importance of understanding system outputs in shaping trust perceptions, yet it often treats trust as a static outcome rather than a dynamic process (Shin, 2021). As a result, existing literature lacks a coherent framework that connects the structural properties of AI systems with the psychological processes through which trust is formed, maintained, or disrupted within organizations. This fragmentation limits the ability of scholars to fully explain why trust breakdown occurs and how it evolves in AI-mediated work environments.

Against this backdrop, a critical gap emerges in the literature. Existing research fails to explain how AI-driven systems structurally and psychologically reshape trust dynamics within organizations. While prior studies acknowledge the importance of trust in technology adoption and organizational functioning, they do not provide a mechanism-based explanation of how specific features of AI systems interact with human cognition, emotions, and social interpretations to produce trust-related outcomes. In particular, there is a lack of integrative models that capture the processes through which AI-induced uncertainty, opacity, and perceived loss of control translate into distrust and resistance. Furthermore, current research does not adequately address how these processes unfold at the organizational level, where individual perceptions aggregate and interact with institutional structures, leadership practices, and cultural norms. The absence of such a framework leaves a significant theoretical and practical void in understanding the challenges associated with AI-driven transformation.

This study aims to address this gap by developing a mechanism-based conceptual framework that explains how AI systems contribute to trust breakdown within organizations. Specifically, this article seeks to elucidate the psychological processes through which the structural characteristics of AI systems influence trust dynamics, integrating insights from artificial intelligence research, organizational behavior, and trust theory. By identifying the mechanisms that link AI-driven work environments to trust-related outcomes, this study contributes to a more comprehensive understanding of the human implications of digital transformation. In doing so, it offers a theoretically grounded and practically relevant perspective on how organizations can better navigate the complexities of trust in the age of artificial intelligence.

2. Literature Review

2.1 Foundations of Organizational Trust

Trust has long been recognized as a fundamental construct in organizational theory, serving as a cornerstone for cooperation, coordination, and effective decision-making. At its core, trust refers to a willingness to accept vulnerability based on positive expectations regarding the intentions or behaviors of another party (Mayer *et al*., 1995). This definition emphasizes two essential elements: the presence of risk and the expectation of favorable outcomes. Trust is not merely a static belief but a dynamic psychological state that enables individuals to engage in actions that expose them to potential harm, uncertainty, or dependence.

A central distinction in the trust literature lies between cognitive and affective dimensions of trust. Cognitive-based trust is grounded in rational assessments of another party's competence, reliability, and integrity, reflecting a calculative evaluation of trustworthiness. In contrast, affective-based trust is rooted in emotional bonds, interpersonal care, and mutual concern, emerging through repeated social interactions and relational experiences (McAllister, 1995). These two dimensions operate simultaneously yet differently, shaping how individuals interpret and respond to uncertainty within organizational contexts. Cognitive trust tends to dominate in formal, task-oriented interactions, while affective trust becomes more salient in relational and long-term exchanges.

The formation of trust is particularly salient in situations characterized by incomplete information and inherent uncertainty. Initial trust, as conceptualized by McKnight *et al.* (1998), often develops in the absence of direct experience, relying on institutional cues, structural assurances, and generalized expectations. This highlights the role of organizational systems and environments in shaping trust perceptions even before interpersonal relationships are fully established. Furthermore, trust is closely linked to behavioral outcomes, as it influences individuals' willingness to engage in risk-taking behaviors that are critical for organizational functioning, such as knowledge sharing, collaboration, and innovation.

Importantly, trust is not only an antecedent to cooperation but also a mechanism that enables action under uncertainty. Individuals who trust are more likely to take risks because they believe that potential negative outcomes will be mitigated or that others will act in their best interest. Empirical research has demonstrated that trust significantly predicts performance-related behaviors and reduces perceived risk in decision-making processes (Colquitt *et al.*, 2007). Thus, trust can be understood as a facilitating condition that allows organizations to function effectively despite uncertainty and complexity. This foundational perspective provides a basis for examining how trust operates and potentially deteriorates in technologically mediated environments.

2.2 Trust in Technology and Automation

While early trust research predominantly focused on interpersonal relationships, subsequent developments have extended the concept to include trust in technological systems and automated agents. As organizations increasingly rely on information systems and automation, trust becomes directed not only toward human actors but also toward technological artifacts. This shift requires a reconceptualization of trust, recognizing that individuals must evaluate the reliability and functionality of systems that do not possess intentions or emotions.

Trust in automation is defined as the attitude that an automated system will help achieve an individual's goals in situations characterized by uncertainty and vulnerability (Lee & See, 2004). This perspective highlights that trust in technology involves similar underlying mechanisms as interpersonal trust, including risk assessment and expectation formation, but differs in its basis, as it relies on perceived system performance, predictability, and transparency rather than social cues. Consequently, individuals develop trust in systems through interactions, feedback, and observed outcomes rather than relational experiences.

A key issue in this domain is the calibration of trust, which refers to the alignment between an individual's level of trust and the actual capabilities of the system. Miscalibrated trust can lead to misuse, disuse, or abuse of technology. Overtrust may result in blind reliance on automated systems, increasing the risk of errors when systems fail, while undertrust may lead to unnecessary rejection of useful technologies (Dzindolet *et al.*, 2003). Achieving appropriate reliance is therefore critical, as it ensures that individuals use technology in a manner consistent with its strengths and limitations.

Further, trust in automation is influenced by multiple factors, including system reliability, feedback mechanisms, and user understanding. Hoff and Bashir (2015) emphasize that trust is shaped not only by system performance but also by user characteristics, environmental context, and prior experiences. These factors interact to determine how individuals interpret system behavior and whether they choose to rely on it. In organizational settings, this complexity is amplified by the integration of multiple technologies and the need for coordination across different levels of decision-making. As a result, trust in technology becomes a multifaceted construct that reflects both individual perceptions and system-level properties.

2.3 Trust in Artificial Intelligence

The emergence of artificial intelligence introduces new challenges and complexities to the concept of trust in technology. Unlike traditional automated systems, AI systems exhibit characteristics such as opacity, autonomy, and unpredictability, which fundamentally alter how trust is formed and maintained. Opacity refers to the difficulty of understanding how AI systems generate their outputs, particularly in the case of complex machine learning models. This lack of transparency limits users' ability to evaluate the reasoning behind decisions, thereby increasing uncertainty.

Autonomy further complicates trust dynamics, as AI systems can operate with minimal human intervention, making decisions that directly impact organizational outcomes. This shift reduces human control and raises concerns about accountability and oversight. At the same time, the unpredictability of AI systems, especially those that learn and adapt over time, challenges traditional assumptions about system reliability and consistency. These characteristics create a unique environment in which trust must be established without full comprehension or control over the system.

Recent studies highlight that trust in AI is influenced not only by technical performance but also by psychological and contextual factors. Afroogh *et al*. (2024) emphasize the evolving nature of trust in AI, noting that users' perceptions are shaped by both system attributes and broader socio-technical contexts. Similarly, Xu *et al*. (2024) demonstrate that leadership and organizational support play a critical role in facilitating initial trust in AI systems, particularly when users lack sufficient understanding of the technology. Cho *et al*. (2025) further show that institutional factors, such as organizational policies and support mechanisms, significantly influence trust in AI adoption.

These findings suggest that trust in AI cannot be fully explained by technological factors alone. Instead, it emerges from the interaction between system characteristics, individual perceptions, and organizational context. This complexity underscores the need for integrative frameworks that capture the multi-level nature of trust in AI-driven environments.

2.4 Explainability, Transparency, and Fairness

Explainability, transparency, and fairness have been widely identified as critical factors influencing trust in AI systems. Explainability refers to the extent to which users can understand the reasoning behind AI-generated decisions. Transparency involves the availability of information about how systems operate, while fairness relates to the perceived equity and impartiality of outcomes. Together, these factors are often assumed to enhance trust by reducing uncertainty and increasing user confidence.

However, empirical evidence suggests that the relationship between these factors and trust is more complex than commonly assumed. While explainability can improve users' understanding of system outputs, it does not always lead to increased trust. In some cases, providing explanations may reveal limitations or biases in the system, thereby reducing confidence. Shin (2021) argues that the effectiveness of explainability depends on how

information is presented and interpreted by users, highlighting the role of cognitive processes in shaping trust perceptions.

The concept of the transparency paradox further illustrates this complexity. Increased transparency does not necessarily lead to greater trust, as users may become overwhelmed by information or may lack the expertise to interpret it effectively. Grimmelikhuijsen (2023) demonstrates that while transparency can enhance perceived trustworthiness, its effects are contingent on users' ability to process and evaluate the information provided. Similarly, Leichtmann *et al*. (2023) find that explainable AI influences trust and behavior differently depending on task context and risk level.

More recent research emphasizes the importance of design and user experience in shaping trust outcomes. Glassberg *et al*. (2025) show that transparency and explainability must be carefully integrated into system design to effectively support trust. These findings suggest that explainability, transparency, and fairness are not sufficient conditions for trust but rather components of a broader socio-technical system that influences how trust is constructed and maintained.

2.5 Algorithm Aversion and Behavioral Response

Despite advancements in AI capabilities, individuals do not consistently trust algorithmic systems. One of the most well-documented phenomena in this area is algorithm aversion, which refers to the tendency of individuals to prefer human judgment over algorithmic recommendations, particularly after observing errors. Dietvorst *et al*. (2015) demonstrate that even minor errors in algorithmic performance can lead to significant reductions in trust, as individuals hold algorithms to higher standards than human decision-makers.

In contrast, research on algorithm appreciation suggests that individuals may prefer algorithmic advice under certain conditions, particularly when tasks are perceived as complex or data-driven (Logg *et al*., 2019). This duality highlights the context-dependent nature of trust in AI, where individuals may oscillate between trust and distrust based on their experiences and perceptions. Hou and Jung (2021) further argue that users' perceptions of expertise play a critical role in determining whether they rely on human or algorithmic advice.

These behavioral responses indicate that trust in AI is not solely determined by system performance but is also influenced by cognitive biases, expectations, and social factors. The sensitivity of trust to perceived errors and inconsistencies suggests that trust in AI is fragile and susceptible to rapid deterioration. This fragility has important implications for organizations, as it can undermine the effectiveness of AI systems and hinder their integration into decision-making processes.

2.6 Trust Under Digital Transformation

The broader context of digital transformation further complicates trust dynamics within organizations. As organizations adopt advanced technologies, employees are exposed to increasing levels of complexity, uncertainty, and workload. Technostress, defined as the stress experienced due to the use of information technologies, has been identified as a significant factor influencing employee behavior and well-being (Ayyagari *et al*., 2011). This stress arises from factors such as information overload, constant connectivity, and the need to adapt to rapidly changing systems.

Tarafdar *et al*. (2015) highlight that technostress can negatively impact performance, reduce job satisfaction, and increase resistance to technological change. These effects are particularly pronounced in environments where employees feel overwhelmed by the pace and complexity of digital transformation. Benbya *et al*. (2021) further emphasize that digital environments are characterized by increasing interdependence and uncertainty, which can exacerbate cognitive and emotional strain.

In such contexts, trust becomes more difficult to establish and maintain. The combination of technological complexity, information overload, and uncertainty creates conditions in which individuals may struggle to evaluate system reliability and make informed decisions. As a result, trust in both technology and organizational structures may erode, leading to disengagement and resistance. Understanding how these contextual factors interact with AI systems is essential for developing a comprehensive explanation of trust dynamics in digitally transformed organizations.

3. Critical Synthesis and Research Gap

3.1 Fragmentation of Existing Literature

The preceding review reveals that the study of trust in contemporary organizations has evolved along multiple disciplinary trajectories, each offering valuable yet partial insights. Research on organizational trust has established a robust theoretical foundation, emphasizing trust as a relational and risk-based construct grounded in perceptions of competence, integrity, and benevolence. In parallel, the literature on trust in technology and automation has extended these ideas to non-human agents, highlighting the importance of system reliability, predictability, and user interaction in shaping trust. More recently, research on artificial intelligence has begun to explore how algorithmic systems influence trust dynamics, particularly through factors such as explainability, transparency, and fairness.

Despite these advances, the literature remains fragmented in its treatment of trust in AI-driven organizational contexts. Studies on artificial intelligence tend to prioritize technological attributes, focusing on system performance, accuracy, and design features, often treating trust as a secondary or outcome variable. In contrast, organizational behavior research emphasizes human cognition, emotion, and behavior but frequently abstracts away from the technological characteristics that increasingly shape workplace interactions. Meanwhile, trust research continues to rely heavily on interpersonal frameworks, which are not fully equipped to account for the unique properties of AI systems, such as autonomy and opacity.

This fragmentation results in a conceptual disconnect. Each stream of research addresses a component of the problem, yet none provides a comprehensive explanation of how these components interact in practice. As organizations integrate AI into core processes, trust is no longer confined to human relationships or isolated technological interactions. Instead, it emerges within complex socio-technical systems where human actors and intelligent technologies are deeply intertwined. The absence of integrative frameworks limits the ability of scholars to capture this complexity and to explain how trust is constructed, challenged, and transformed in AI-mediated environments.

3.2 Missing Mechanistic Link Between AI Systems and Trust Dynamics

A critical limitation of existing research lies in its insufficient attention to the mechanisms through which AI systems influence trust. Much of the literature adopts a correlational perspective, identifying factors that are associated with higher or lower levels of trust without explicating the underlying processes that drive these relationships. For instance, studies on explainability and transparency demonstrate that these features can affect trust perceptions, yet they often do not clarify how individuals cognitively and emotionally process such information to arrive at trust-related judgments.

Similarly, research on algorithm aversion and reliance highlights behavioral responses to AI systems but does not fully explain the psychological pathways that lead to these responses. The tendency to distrust algorithms after observing errors suggests the presence of deeper cognitive and affective processes, such as heightened sensitivity to perceived uncertainty or

violations of expectations. However, these processes are rarely modeled explicitly within existing frameworks.

This gap becomes more pronounced when considering the structural characteristics of AI systems. Features such as opacity, autonomy, and unpredictability are frequently cited as sources of distrust, yet their effects are typically inferred rather than systematically theorized. There is limited understanding of how these features interact with individual perceptions of control, fairness, and identity to produce trust-related outcomes. Without a mechanism-based explanation, it is difficult to determine whether trust breakdown is driven primarily by system design, user interpretation, or contextual factors within the organization.

Furthermore, existing models often treat trust as a static construct, overlooking its dynamic and processual nature. In reality, trust evolves over time, shaped by ongoing interactions between individuals and systems, as well as by feedback from organizational environments. The absence of temporal and process-oriented perspectives restricts the ability to capture how trust is gradually built, disrupted, or recalibrated in response to AI integration.

3.3 3.3. Toward a Mechanism-Based Understanding of Trust Breakdown

Taken together, these limitations point to a fundamental gap in the literature: the lack of a coherent, mechanism-based framework that links AI-driven system characteristics to trust dynamics within organizations. Trust breakdown in AI contexts cannot be adequately explained by isolated factors or single-level analyses. Instead, it reflects a multi-layered process in which structural features of technology interact with psychological responses and organizational conditions.

This study argues that trust breakdown is not a direct or automatic consequence of AI adoption. Rather, it is the outcome of unaddressed psychological mechanisms that are triggered when individuals engage with AI-driven systems under conditions of uncertainty, limited transparency, and reduced control. These mechanisms may include cognitive ambiguity, perceived loss of agency, fairness concerns, and threats to professional identity. When these processes are activated without adequate organizational support or design interventions, they can undermine trust and lead to behavioral responses such as skepticism, resistance, and disengagement.

By shifting the focus from outcomes to processes, a mechanism-based perspective offers a more nuanced understanding of trust in AI-driven organizations. It enables the identification of specific pathways through which trust is constructed and disrupted, providing a foundation for both theoretical advancement and practical intervention. Such an approach also facilitates the integration of previously fragmented research streams, as it explicitly connects technological characteristics, psychological processes, and organizational contexts within a unified framework.

3.4 Research Gap and Direction for Conceptual Development

Building on this synthesis, a clear research gap emerges. Existing research fails to explain how AI-driven systems structurally and psychologically reshape trust dynamics within organizations. While prior studies have identified relevant factors and outcomes, they have not articulated the mechanisms that connect these elements in a systematic and integrative manner. In particular, there is a lack of conceptual models that explain how the structural properties of AI systems translate into psychological responses and, ultimately, into trust-related behaviors at the organizational level.

Addressing this gap requires a shift toward theory development that is explicitly mechanism-oriented and multi-level in nature. Such an approach must account for the interplay between system design, individual cognition and emotion, and organizational structures. It must also

recognize that trust in AI is not merely an extension of traditional trust constructs but a distinct phenomenon shaped by the unique characteristics of intelligent systems.

Accordingly, the next section develops a conceptual framework that seeks to explain how AI system characteristics trigger specific psychological mechanisms that lead to trust breakdown in organizational settings. By articulating these relationships, the study aims to provide a more comprehensive and actionable understanding of trust in the context of AI-driven transformation.

4. Conceptual Framework

4.1 Overview of the Conceptual Model

Building on the identified research gap, this study proposes a mechanism-based conceptual framework that explains how trust breakdown emerges in AI-driven organizations. The model is structured around a causal sequence in which the structural characteristics of AI systems shape individual psychological responses, which in turn lead to trust-related outcomes. Specifically, the framework posits that AI system characteristics act as exogenous drivers, triggering psychological mechanisms that mediate the relationship between technology and trust, ultimately resulting in varying degrees of trust breakdown within organizational contexts.

This visual articulates the mechanism-based architecture through which AI system characteristics translate into trust breakdown within organizations. It emphasizes the sequential logic linking structural properties, psychological processes, and behavioral outcomes, while embedding organizational conditions as contextual moderators.

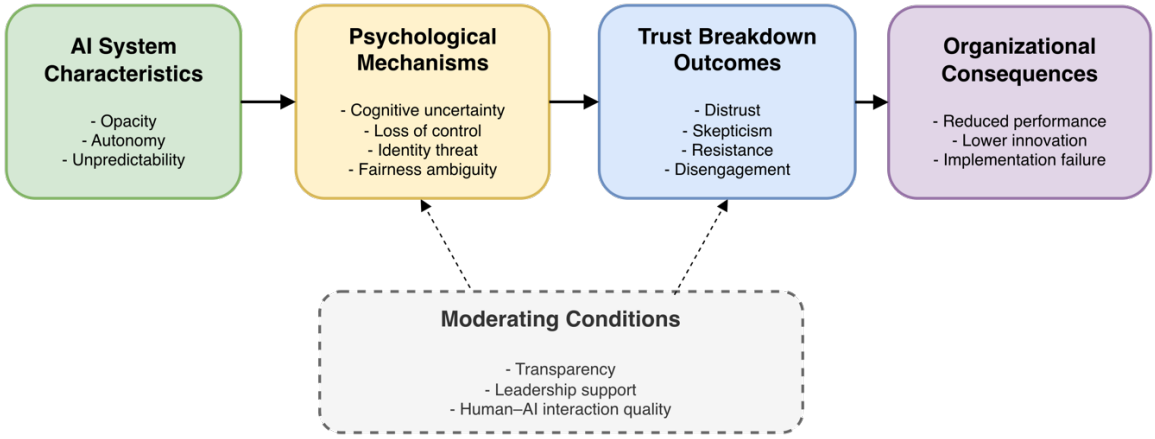

Figure 1. Mechanism-Based Model of Trust Breakdown in AI-Driven Organizations
Source: Developed by the author

As illustrated in Figure 1, the model specifies a sequential mechanism in which structural characteristics of AI systems initiate psychological responses that ultimately manifest as trust-related behavioral outcomes. Figure 1 clarifies that trust breakdown is not a direct consequence of AI adoption but emerges through mediated processes, while organizational conditions shape the strength and direction of these relationships. The framework thereby integrates technological, psychological, and organizational dimensions into a coherent causal architecture that underpins the article's core argument.

This perspective departs from prior approaches that treat trust as a direct function of system attributes or user attitudes. Instead, it emphasizes that trust is not formed or eroded in isolation but is the product of an interpretive process through which individuals make sense of AI systems. By focusing on these intervening mechanisms, the model captures the

dynamic and processual nature of trust, allowing for a more nuanced understanding of how trust is constructed, destabilized, and potentially recalibrated in AI-mediated environments.

4.2 AI System Characteristics as Structural Drivers

The first component of the model concerns the structural properties of AI systems that shape how individuals perceive and interact with them. Three characteristics are particularly salient in influencing trust dynamics: opacity, autonomy, and unpredictability.

This table consolidates the core constructs used in the conceptual model, ensuring definitional clarity and theoretical precision. It reduces ambiguity by aligning each construct

Table 1. Key Constructs and Conceptual Definitions

Construct	Definition	Role in the Model
AI System Characteristics	Structural properties of AI systems, including opacity, autonomy, and unpredictability, that shape how decisions are generated and perceived	Exogenous drivers that initiate the trust formation or breakdown process
Cognitive Uncertainty	A state in which individuals are unable to fully understand or evaluate AI-generated decisions due to limited interpretability	Psychological mechanism that undermines confidence in system reliability
Perceived Loss of Control	The perception that one's ability to influence decisions is reduced due to AI autonomy	Psychological mechanism that decreases agency and increases resistance
Identity Threat	A perceived challenge to one's professional role, expertise, or self-concept due to AI capabilities	Psychological mechanism that triggers defensive reactions toward AI
Fairness Ambiguity	Uncertainty regarding whether AI-generated outcomes are equitable, unbiased, or justified	Psychological mechanism that weakens perceived legitimacy of AI systems
Trust Breakdown	A decline in confidence, acceptance, and willingness to rely on AI systems	Aggregate outcome reflecting erosion of trust
Behavioral Responses	Observable reactions such as distrust, skepticism, resistance, and disengagement	Manifestation of trust breakdown at the behavioral level
Organizational Conditions	Contextual factors such as transparency, leadership support, and human–AI interaction quality	Moderators that influence the strength and direction of relationships

Source: Developed by the authors

Table 1 clarifies the conceptual architecture by explicitly defining each construct and situating it within the model's causal logic. By distinguishing between drivers, mechanisms, outcomes, and moderators, Table 1 supports analytical precision and helps prevent construct ambiguity, thereby strengthening the theoretical coherence of the framework.

Opacity refers to the limited transparency of AI decision-making processes, especially in complex machine learning models where the underlying logic is not readily interpretable. When users are unable to understand how decisions are generated, they face increased cognitive uncertainty, which complicates their ability to evaluate the system's reliability and fairness. This lack of interpretability reduces the basis for informed trust and may lead to skepticism or hesitation in relying on AI outputs (Shin, 2021; Zerilli *et al.*, 2022).

Autonomy reflects the extent to which AI systems operate independently of human intervention. As AI systems assume greater decision-making authority, the locus of control shifts away from human actors, potentially diminishing individuals' sense of agency. While autonomy can enhance efficiency, it also raises concerns about accountability and oversight, particularly when decisions have significant organizational consequences. This shift can

create tension between perceived system competence and reduced human control, affecting trust formation.

Unpredictability arises from the adaptive and probabilistic nature of many AI systems, which may produce outputs that vary across contexts or over time. Unlike deterministic systems, AI models can evolve through learning processes, making their behavior less consistent and harder to anticipate. This variability challenges users' expectations of stability and reliability, increasing perceived risk in decision-making situations. Together, these characteristics create a structural environment that is inherently conducive to uncertainty and interpretive ambiguity.

4.3 Psychological Mechanisms as Mediating Processes

Central to the proposed framework is the role of psychological mechanisms that mediate the relationship between AI system characteristics and trust outcomes. These mechanisms represent the internal processes through which individuals interpret, evaluate, and respond to AI-driven environments. Four key mechanisms are identified: cognitive uncertainty, perceived loss of control, identity threat, and fairness ambiguity.

Cognitive uncertainty emerges when individuals are unable to fully comprehend how AI systems generate their outputs. This uncertainty is amplified by system opacity and complexity, leading to difficulties in assessing the validity of decisions. When individuals lack sufficient understanding, they may rely on heuristics or default to skepticism, particularly in high-stakes contexts. As a result, cognitive uncertainty undermines the confidence required to establish trust.

Perceived loss of control reflects the extent to which individuals feel that their ability to influence outcomes is diminished by the presence of AI systems. As decision-making authority is transferred to algorithms, individuals may experience a reduction in autonomy and agency. This perception can generate discomfort and resistance, as control is a fundamental component of psychological security in organizational settings. When individuals perceive that they are no longer active participants in decision processes, their willingness to trust the system may decline.

Identity threat arises when AI systems challenge individuals' professional roles, expertise, or sense of self. In knowledge-intensive environments, employees often derive identity from their cognitive contributions and decision-making capabilities. The introduction of AI systems that perform similar or superior functions can create a perceived threat to this identity, leading to defensive responses. Such responses may include rejecting or undermining AI outputs, not necessarily due to system deficiencies but as a means of preserving self-concept.

Fairness ambiguity refers to uncertainty regarding the equity and impartiality of AI-generated decisions. When individuals cannot determine whether outcomes are fair, unbiased, or justified, they may question the legitimacy of the system. This ambiguity is particularly problematic in contexts where decisions have distributive or evaluative consequences, such as performance assessments or resource allocation. Perceptions of unfairness, even if not empirically substantiated, can significantly erode trust.

These psychological mechanisms are not mutually exclusive but interact in complex ways. For example, cognitive uncertainty may amplify perceptions of unfairness, while loss of control may intensify identity threat. The combined effect of these mechanisms creates a reinforcing cycle that increases the likelihood of trust breakdown.

This visual isolates and decomposes the internal dynamics of psychological mechanisms, clarifying how they interact and reinforce one another in producing trust breakdown. It shifts the analytical focus from linear causality to interdependent cognitive–emotional processes operating within individuals.

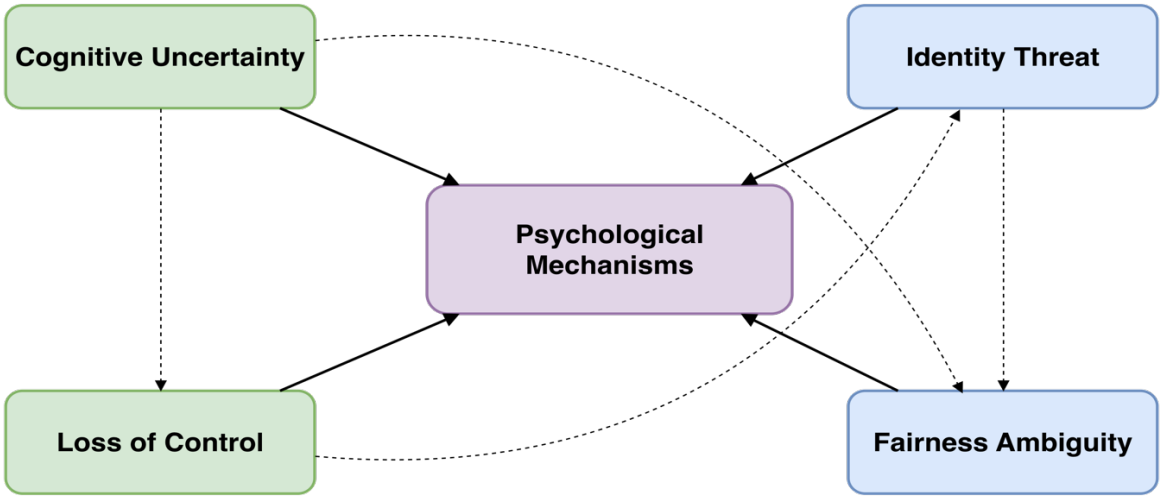

Figure 2. Interdependent Psychological Mechanisms Underlying Trust Breakdown
Source: Developed by the author

The structure articulated in Figure 2 isolates the internal configuration of psychological mechanisms and shows that they do not operate independently but form a mutually reinforcing system. Figure 2 clarifies how cognitive uncertainty can amplify perceptions of unfairness, while loss of control intensifies identity threat, generating compounding effects that accelerate trust erosion. By making these interdependencies explicit, the figure strengthens the article's mechanism-based argument and demonstrates that trust breakdown emerges from interacting psychological processes rather than isolated responses.

4.4 Trust Breakdown as Behavioral and Attitudinal Outcomes

The activation of these psychological mechanisms leads to a range of trust-related outcomes that collectively constitute trust breakdown. Rather than viewing trust as a binary state, the framework conceptualizes trust breakdown as a continuum of declining confidence, reliance, and acceptance of AI systems within organizational settings.

This table specifies the precise relationships between psychological mechanisms and trust-related outcomes, making explicit the pathways that are only implicitly described in the text. It enhances analytical precision by preventing overly generalized or ambiguous causal claims.

Table 2. Mapping Psychological Mechanisms to Trust-Related Outcomes

Psychological Mechanism	Primary Outcome(s)	Nature of Effect
Cognitive Uncertainty	Skepticism; Distrust	Reduces confidence in decision validity and increases reliance on doubt-based evaluation
Perceived Loss of Control	Resistance; Disengagement	Triggers withdrawal and opposition due to diminished agency
Identity Threat	Resistance; Distrust	Generates defensive responses aimed at protecting professional self-concept
Fairness Ambiguity	Distrust; Skepticism	Undermines perceived legitimacy and fairness of AI-generated outcomes
Combined Mechanisms	Disengagement; Systemic Trust Erosion	Reinforcing effects that amplify and accelerate overall trust breakdown

Source: Developed by the authors

Table 2 sharpens the model's explanatory power by linking each psychological mechanism to specific behavioral and attitudinal consequences. Rather than treating trust breakdown as a uniform outcome, Table 2 demonstrates that different mechanisms produce distinct

patterns of response, thereby supporting a more nuanced and mechanism-driven interpretation of trust dynamics in AI-driven organizations.

One primary outcome is distrust, characterized by negative expectations regarding the system's intentions or performance. Distrust goes beyond the absence of trust and reflects an active belief that the system may produce undesirable or harmful outcomes. This can lead to heightened scrutiny and rejection of AI-generated decisions.

Skepticism represents a more moderate form of trust erosion, where individuals question the validity of system outputs without fully rejecting them. Skeptical users may engage in additional verification or seek alternative sources of information, reducing the efficiency gains associated with AI adoption.

Resistance to AI systems is another critical outcome, manifesting in behaviors such as avoidance, non-compliance, or active opposition to system implementation. Resistance may stem from perceived threats to control, identity, or fairness and can significantly hinder organizational transformation efforts.

Finally, disengagement reflects a broader withdrawal from interaction with AI systems and, in some cases, from organizational processes more generally. When individuals perceive that systems are unreliable or misaligned with their values, they may reduce their level of involvement, leading to decreased performance and innovation.

These outcomes illustrate that trust breakdown has both attitudinal and behavioral dimensions, affecting not only how individuals perceive AI systems but also how they act within organizational contexts.

4.5 Moderating Conditions in Organizational Contexts

While the framework emphasizes the role of structural and psychological factors, it also recognizes that the relationship between AI systems and trust outcomes is contingent on organizational conditions. Several moderating factors can influence the extent to which psychological mechanisms lead to trust breakdown.

Organizational transparency plays a critical role in shaping how individuals interpret AI systems. Clear communication about how systems function, why decisions are made, and what limitations exist can reduce cognitive uncertainty and fairness ambiguity. However, as noted earlier, transparency must be carefully designed to be meaningful and accessible to users.

Leadership support is another important moderator. Leaders who actively endorse AI systems, provide guidance, and address employee concerns can facilitate trust formation by signaling institutional commitment and reducing perceived risks. Leadership behavior can also influence how employees interpret system-related changes, framing them as opportunities rather than threats (Xu *et al.*, 2024; Cho *et al.*, 2025).

The quality of human–AI interaction further moderates trust dynamics. Systems that are designed with user experience in mind, including intuitive interfaces and effective feedback mechanisms, are more likely to support positive trust perceptions. Conversely, poorly designed interactions can exacerbate cognitive and emotional strain, reinforcing negative psychological responses (Cui *et al.*, 2025).

These moderating factors highlight that trust in AI-driven organizations is not determined solely by technology or individual psychology but is shaped by the broader organizational environment. Effective management of these conditions can mitigate the negative effects of AI system characteristics and support more stable trust relationships.

4.6 Summary of the Conceptual Framework

In summary, the proposed framework conceptualizes trust breakdown in AI-driven organizations as a multi-stage process in which structural characteristics of AI systems trigger psychological mechanisms that lead to attitudinal and behavioral outcomes. By incorporating moderating organizational conditions, the model provides a comprehensive and integrative explanation of trust dynamics in contemporary work environments.

This mechanism-based perspective advances the literature by moving beyond static and fragmented approaches, offering a structured understanding of how trust is disrupted in the context of AI adoption. It also lays the foundation for future empirical research and practical interventions aimed at designing AI systems and organizational practices that support, rather than undermine, trust.

5. Discussion

5.1 Theoretical Implications

The proposed framework contributes to the literature by advancing a mechanism-based understanding of trust in AI-driven organizations, addressing a critical gap identified in prior research. Existing studies have largely treated trust as either an outcome of technological attributes or as a relational construct rooted in interpersonal interactions. By integrating these perspectives, this study reconceptualizes trust as a process that emerges from the interaction between structural system characteristics and human psychological responses. This shift from a static to a process-oriented view enables a more precise explanation of how trust is formed, destabilized, and recalibrated in complex socio-technical environments.

First, this study extends traditional organizational trust theory by incorporating non-human agents as central actors in trust dynamics. Classical models emphasize perceptions of ability, benevolence, and integrity as the foundation of trust (Mayer *et al.*, 1995), yet these dimensions are inherently human-centric. In AI-driven contexts, however, individuals must evaluate systems that lack intentionality, emotion, and moral agency. The framework demonstrates that trust in such systems is mediated not only by perceptions of performance but also by psychological mechanisms such as uncertainty, perceived loss of control, and identity threat. This insight suggests that trust theory must evolve to account for the unique epistemic and ontological characteristics of AI systems.

Second, the study contributes to the literature on trust in technology by clarifying the role of psychological mediation. While prior research has emphasized the importance of system reliability, transparency, and usability (Lee & See, 2004; Hoff & Bashir, 2015), it has often overlooked the internal processes through which individuals interpret these attributes. By identifying specific psychological mechanisms, the framework provides a more granular understanding of how technological features translate into trust-related outcomes. This contribution is particularly important in AI contexts, where system complexity and opacity amplify the role of subjective interpretation.

Third, this study bridges the gap between artificial intelligence research and organizational behavior by situating AI within a broader behavioral and organizational context. Much of the AI literature focuses on technical performance and optimization, whereas organizational behavior research emphasizes human attitudes and behaviors without fully accounting for technological mediation. The proposed framework integrates these perspectives by showing how AI system characteristics interact with human cognition and organizational conditions to shape trust dynamics. This integrative approach responds to calls for interdisciplinary research that captures the socio-technical nature of contemporary organizations.

Finally, the framework contributes to the emerging discourse on explainability and fairness in AI by highlighting their conditional effects on trust. Rather than assuming a linear relationship between transparency and trust, the study demonstrates that these factors

operate through psychological mechanisms that may produce unintended consequences. For instance, increased transparency may reduce uncertainty but simultaneously expose system limitations, thereby undermining confidence. This nuanced perspective challenges simplistic assumptions and encourages more sophisticated theorization of how design features influence user perceptions.

5.2 Practical Implications

The findings of this study have important implications for organizations seeking to implement AI systems effectively. A key insight is that trust in AI cannot be achieved solely through technological enhancement. Improving algorithmic accuracy or efficiency, while necessary, is insufficient to ensure trust if underlying psychological mechanisms are not addressed. Organizations must therefore adopt a more holistic approach that considers both system design and human experience.

One critical implication is the need to redesign organizational approaches to AI implementation with a focus on trust as a central objective. Rather than treating trust as an emergent byproduct, organizations should explicitly design for trust by addressing sources of cognitive uncertainty, perceived loss of control, and fairness ambiguity. This may involve providing clear and accessible explanations of AI decisions, establishing feedback mechanisms that allow users to question or override system outputs, and ensuring that AI systems align with organizational values and norms.

Another important implication concerns the role of leadership. Leaders play a pivotal role in shaping how employees interpret and respond to AI systems. By communicating the purpose, benefits, and limitations of AI, leaders can reduce uncertainty and frame technological change in a more positive light. Moreover, leadership behaviors that emphasize support, inclusivity, and transparency can help mitigate identity threats and foster a sense of psychological safety. This suggests that successful AI adoption requires not only technical expertise but also effective change management and leadership practices.

The design of human–AI interaction also emerges as a critical factor. Systems that are intuitive, responsive, and aligned with user needs are more likely to support trust formation. Conversely, poorly designed interfaces that obscure system logic or limit user control can exacerbate negative psychological responses. Organizations should therefore invest in user-centered design principles that prioritize usability, clarity, and interaction quality. This includes considering how information is presented, how users can engage with the system, and how feedback is incorporated into system behavior.

Finally, organizations must recognize the broader contextual factors that influence trust. Digital transformation often introduces increased complexity, workload, and uncertainty, which can amplify technostress and reduce employees' capacity to engage with AI systems effectively. Addressing these issues requires organizational interventions that go beyond the technology itself, such as workload management, training programs, and support systems that help employees adapt to new ways of working. By creating an environment that supports both technological and human adaptation, organizations can better sustain trust in AI-driven contexts.

5.3 Managerial Insights

From a managerial perspective, the central implication of this study is that trust in AI is not an automatic outcome of technological advancement but a condition that must be actively cultivated. Managers often assume that employees will accept AI systems if they demonstrate superior performance. However, the framework suggests that performance alone is insufficient to secure trust, particularly when systems are perceived as opaque, autonomous, or unpredictable.

Managers should therefore shift their focus from solely optimizing system performance to managing the human experience of interacting with AI. This involves recognizing that employees' perceptions of control, fairness, and identity are critical determinants of trust. For instance, allowing employees to maintain a degree of control over decision processes can reduce resistance and increase acceptance. Similarly, ensuring that AI systems are perceived as fair and aligned with organizational values can strengthen legitimacy and trust.

Another key insight is the importance of trust calibration. Managers must ensure that employees neither overtrust nor undertrust AI systems. Overtrust can lead to blind reliance and potential errors, while undertrust can result in underutilization of valuable technologies. Achieving appropriate reliance requires continuous monitoring, feedback, and adjustment, as well as clear communication about system capabilities and limitations.

In addition, managers should view trust as a collective and organizational phenomenon rather than an individual-level issue. Trust breakdown can spread across teams and departments, affecting overall organizational performance. As such, interventions aimed at building trust should be coordinated and systemic, involving multiple stakeholders and organizational levels.

Ultimately, the findings suggest that effective AI adoption depends on the ability of organizations to align technological capabilities with human needs and expectations. Managers who recognize and address the psychological dimensions of trust are more likely to succeed in integrating AI systems into their organizations, while those who neglect these aspects risk encountering resistance, disengagement, and suboptimal outcomes.

6. Conclusion

This study set out to address a critical gap in understanding how trust operates and deteriorates within AI-driven organizational contexts. While artificial intelligence has been widely recognized as a transformative force capable of enhancing efficiency, accuracy, and decision quality, its integration into organizational systems introduces new and complex challenges related to trust. Existing research has provided valuable insights into trust in interpersonal relationships, technological systems, and algorithmic decision-making. However, these perspectives have remained largely fragmented, limiting their ability to explain how trust dynamics evolve in environments where human and intelligent systems are deeply intertwined.

By developing a mechanism-based conceptual framework, this study offers a more integrated and process-oriented understanding of trust breakdown in AI-driven organizations. The framework demonstrates that trust erosion is not a direct or inevitable consequence of AI adoption. Rather, it emerges through a series of psychological mechanisms triggered by the structural characteristics of AI systems, including opacity, autonomy, and unpredictability. These characteristics give rise to cognitive uncertainty, perceived loss of control, identity threat, and fairness ambiguity, which collectively shape individuals' interpretations of and responses to AI systems. As these mechanisms intensify, they can lead to distrust, skepticism, resistance, and disengagement, ultimately undermining the effectiveness of AI implementation.

A key contribution of this study lies in shifting the focus from static conceptualizations of trust to a dynamic, mechanism-based perspective. Trust is not simply a condition that exists or does not exist; it is a process that unfolds through ongoing interactions between individuals, technologies, and organizational structures. This perspective enables a deeper understanding of why trust breakdown occurs and highlights the importance of addressing the underlying psychological processes rather than focusing solely on technological performance or system design.

The study also underscores the importance of organizational context in shaping trust dynamics. Factors such as transparency, leadership support, and the quality of human–AI interaction can significantly influence how individuals interpret and respond to AI systems. These moderating conditions suggest that trust in AI is not determined solely by the technology itself but is co-constructed within organizational environments. As such, organizations must adopt a holistic approach to AI implementation that integrates technological, psychological, and organizational considerations.

From a broader perspective, the findings highlight a fundamental challenge facing modern organizations: the need to align rapidly advancing technological systems with the inherently bounded capacities of human cognition, emotion, and social interpretation. Failure to address this alignment may result in persistent trust deficits, limiting the potential benefits of AI and hindering organizational performance. Conversely, organizations that successfully manage the interplay between technology and human behavior are more likely to build sustainable trust and realize the full value of AI-driven transformation.

In conclusion, trust in AI-driven organizations should be understood not as a technological outcome but as a psychologically mediated organizational phenomenon. Advancing this understanding is essential for both theory and practice, as it provides a foundation for developing more effective strategies to design, implement, and manage AI systems in ways that support, rather than undermine, trust.

References

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications, 11*, 1568. https://doi.org/10.1057/s41599-024-04044-8

Ayyagari, R., Grover, V., & Purvis, R. (2011). Technostress: Technological antecedents and implications. *MIS Quarterly, 35*(4), 831–858. https://doi.org/10.2307/41409963

Benbya, H., Nan, N., Tanriverdi, H., & Yoo, Y. (2021). Complexity and information systems research in the emerging digital world. *MIS Quarterly, 45*(1), 1–17. https://doi.org/10.25300/MISQ/2021/14915

Benda, N. C., Novak, L. L., Reale, C., & Ancker, J. S. (2022). Trust in AI: Why we should be designing for appropriate reliance. *Journal of the American Medical Informatics Association, 29*(1), 207–212. https://doi.org/10.1093/jamia/ocab238

Brown, A. S., Dishop, C. R., Kuznetsov, A., Chao, P.-Y., & Woolley, A. W. (2025). Beyond efficiency: Trust, AI, and surprise in knowledge work environments. *Computers in Human Behavior, 167*, 108605. https://doi.org/10.1016/j.chb.2025.108605

Cho, S., Hur, J.-Y., & Kim, D. (2025). Bridging trust in AI and its adoption: The role of organizational support in AI chatbot implementation in Korean government agencies. *Government Information Quarterly, 42*(4), 102081. https://doi.org/10.1016/j.giq.2025.102081

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*(4), 909–927. https://doi.org/10.1037/0021-9010.92.4.909

Cui, M., Li, W., & Kamoche, K. (2025). Building trust in decision-support artificial intelligence: A boundary spanning perspective. *Information Systems Journal*. Advance online publication. https://doi.org/10.1111/isj.70015

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dirks, K. T., & Ferrin, D. L. (2001). The role of trust in organizational settings. *Organization Science, 12*(4), 450–467. https://doi.org/10.1287/orsc.12.4.450.10640

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. https://doi.org/10.1016/S1071-5819(03)00038-7

Glassberg, I., Brender-Ilan, Y., & Zwilling, M. (2025). The key role of design and transparency in enhancing trust in AI-powered digital agents. *Journal of Innovation & Knowledge, 10*(5), 100770. https://doi.org/10.1016/j.jik.2025.100770

Grimmelikhuijsen, S. (2023). Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Administration Review, 83*(2), 241–262. https://doi.org/10.1111/puar.13483

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hou, Y. T.-Y., & Jung, M. F. (2021). Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), Article 477. https://doi.org/10.1145/3479864

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

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior, 139*, 107539. https://doi.org/10.1016/j.chb.2022.107539

Liu, C., Liao, Q., & Lu, J. (2026). Employees' trust in AI and innovative behavior: A JD-R model perspective. *Behavioral Sciences, 16*(3), 425. https://doi.org/10.3390/bs16030425

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. https://doi.org/10.2307/258792

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal, 38*(1), 24–59. https://doi.org/10.5465/256727

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. https://doi.org/10.5465/amr.1998.926622

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23*(3), 393–404. https://doi.org/10.5465/amr.1998.926617

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Tarafdar, M., Pullins, E. B., & Ragu-Nathan, T. S. (2015). Technostress: Negative effect on performance and possible mitigations. *Information Systems Journal, 25*(2), 103–132. https://doi.org/10.1111/isj.12042

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems, 28*(2), 118–144. https://doi.org/10.1016/j.jsis.2019.01.003

Xu, Y., Huang, Y., Wang, J., & Zhou, D. (2024). How do employees form initial trust in artificial intelligence: Hard to explain but leaders help. *Asia Pacific Journal of Human Resources, 62*(3), e12402. https://doi.org/10.1111/1744-7941.12402

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns, 3*(4), 100455. https://doi.org/10.1016/j.patter.2022.100455