Manavest
Journal of Human Capital and Organizational Behavior

Vol. 1 | Issue 1 | April 2026
Pages 102-119

Human–AI Misalignment: A Mechanism-Based Framework for Socio-Technical Breakdown

Richard Abdulloh Mundzir[1]
Ahmad Mundzir[2]*

***Corresponding Mail:**
richard.mundzir@udex.or.id

Article History:
Submitted: 23-12-2025
Approved: 16-03-2026
Published: 06-04-2026

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

Abstrak

This article examines the persistent challenges in human–AI collaboration, where advances in artificial intelligence improve decision accuracy yet frequently fail to produce effective socio-technical integration. Existing literature offers fragmented insights into trust, behavior, and system performance, but lacks a coherent, mechanism-based explanation of why collaboration breakdowns occur. To address this gap, the study aims to develop an integrative conceptual framework that explains how discrepancies between AI systems and human actors generate dysfunctional outcomes. Adopting a socio-technical and mechanism-driven analytical approach, the article introduces the concept of Human–AI Misalignment as a multidimensional construct encompassing cognitive, emotional, agency-related, and meaning-based discrepancies. The framework identifies key antecedents such as algorithmic opacity, task complexity, and AI autonomy, which activate psychological mechanisms including cognitive overload, identity threat, perceived loss of control, and perceived unfairness, ultimately shaping behavioral responses such as algorithm aversion, blind reliance, resistance, and disengagement. The study contributes theoretically by reframing AI integration from adoption to alignment, offering a unified explanation of socio-technical breakdowns, and providing a foundation for future empirical research and human-centered AI system design.

Keywords

human–AI misalignment; socio-technical systems; algorithmic decision-making; trust in AI; behavioral responses to AI; organizational performance

[1] Department of Communication Studies, Universitas Terbuka, Jakarta, Indonesia
[2] VISTA Research Center, UDEX Institute, Tasikmalaya, Indonesia

1. Introduction

Artificial intelligence (AI) has rapidly evolved from a peripheral technological tool into a central driver of organizational transformation. Across industries, AI systems are increasingly embedded in decision-making, coordination, and performance management processes, fundamentally reshaping how work is designed and executed (Davenport et al., 2020; Berente et al., 2021; Huang & Rust, 2021). Rather than simply automating routine tasks, AI is now actively involved in generating insights, guiding strategic decisions, and reallocating authority within organizations (Faraj et al., 2018; Murray et al., 2021). This shift reflects a broader transition from technology as support to technology as an embedded socio-technical actor, capable of influencing both organizational structures and epistemic processes (von Krogh, 2018; Shrestha et al., 2019).

Despite these advancements, the integration of AI has produced a persistent paradox. While AI promises improved efficiency, accuracy, and scalability, empirical evidence consistently shows that its adoption does not automatically translate into effective human–AI collaboration (Raisch & Krakowski, 2021; Tambe et al., 2019). Instead, organizations frequently encounter resistance, distrust, and inconsistent behavioral responses when individuals interact with AI systems (Dietvorst et al., 2015; Longoni et al., 2019; Logg et al., 2019). Users may reject algorithmic recommendations even when they outperform human judgment, or alternatively, rely on them uncritically without sufficient evaluation (Prahl & Van Swol, 2017; Burton et al., 2020). These contradictory responses highlight a fundamental limitation in current understandings of AI integration, suggesting that technological capability alone is insufficient to ensure effective collaboration.

This paradox becomes more pronounced when considering the distinct logics underlying human and artificial intelligence. AI systems operate through probabilistic modeling, large-scale data processing, and optimization functions, producing outputs that prioritize statistical accuracy (Burrell, 2016; Agrawal et al., 2019). In contrast, human cognition is shaped by bounded rationality, heuristics, and contextual interpretation, often privileging interpretability and coherence over optimality (Jarrahi, 2018; Castelo et al., 2019). These differences create epistemic and cognitive tensions, where algorithmic outputs may be technically valid but perceived as opaque or misaligned with human reasoning (Yeomans et al., 2019; Arrieta et al., 2020). As a result, interactions between humans and AI frequently generate friction, leading to what appears as irrational or inconsistent behavioral responses.

Beyond cognition, the challenges of human–AI interaction extend into emotional and social domains. AI systems lack emotional awareness and relational sensitivity, yet human decision-making is deeply influenced by affective responses and social meaning (Huang & Rust, 2018; Puntoni et al., 2021). This mismatch can lead to discomfort, anxiety, and resistance, particularly in contexts requiring empathy or moral judgment (Granulo et al., 2019; Araujo et al., 2020). Moreover, the increasing autonomy of AI systems introduces ambiguity in control and accountability, challenging traditional notions of agency within organizations (Murray et al., 2021; Shrestha et al., 2019). Individuals may struggle to calibrate their reliance on AI, oscillating between over-dependence and rejection, both of which undermine decision quality.

Existing literature has begun to explore these issues across multiple streams. Research on algorithm aversion and appreciation highlights the conditional nature of trust in AI systems (Dietvorst et al., 2015; Logg et al., 2019). Studies on socio-technical systems emphasize the importance of aligning technological capabilities with human practices and organizational contexts (Cascio & Montealegre, 2016; Seeber et al., 2020). Meanwhile, trust research underscores the role of transparency, fairness, and explainability in shaping human responses to AI (Glikson & Woolley, 2020; Hoff & Bashir, 2015). Although these perspectives offer valuable insights, they remain largely fragmented, focusing on isolated dimensions of human–AI interaction without providing an integrated explanation of how and why collaboration breakdowns occur.

A critical limitation of prior research lies in its emphasis on outcomes rather than underlying processes. Much of the literature examines phenomena such as trust, resistance, or performance in isolation, without fully theorizing the mechanisms that connect AI characteristics to human behavioral responses (Berente et al., 2021; Lebovitz et al., 2022). Furthermore, AI is often treated as a neutral or external factor, rather than as an active component within a socio-technical system that co-shapes cognition, emotion, and organizational dynamics (Faraj et al., 2018; Murray et al., 2021). This has resulted in a lack of mechanism-based explanations that account for the systemic nature of human–AI interaction.

To address this gap, this paper introduces the concept of Human–AI Misalignment, defined as a systematic discrepancy between the operational logic of AI systems and the cognitive, emotional, and social structures of human actors, resulting in dysfunctional collaboration outcomes. Rather than viewing resistance or misuse of AI as isolated behavioral anomalies, this study conceptualizes such responses as manifestations of deeper misalignments across multiple dimensions. Specifically, we propose that misalignment emerges through four interrelated dimensions: cognitive, emotional, agency-related, and meaning-based discrepancies between humans and AI.

Building on socio-technical systems theory and recent advances in AI research, this paper develops a mechanism-based framework that explains how misalignment arises and translates into observable organizational outcomes. We identify key antecedents, including algorithmic opacity, task complexity, and AI autonomy, which trigger psychological mechanisms such as cognitive overload, identity threat, perceived loss of control, and perceived unfairness. These mechanisms, in turn, shape behavioral responses such as algorithm aversion, blind reliance, resistance, and disengagement, ultimately influencing organizational performance (Endsley, 2017; Binns et al., 2018). Importantly, we also highlight the role of moderating factors, including trust, psychological safety, and AI literacy, in shaping the extent to which misalignment affects outcomes (Glikson & Woolley, 2020; Haque et al., 2023).

This study makes three primary contributions. First, it introduces Human–AI Misalignment as a novel construct that captures the multidimensional nature of breakdowns in human–AI collaboration. Second, it advances a mechanism-based explanation that links AI characteristics to psychological processes and behavioral outcomes, addressing a key gap in existing research. Third, it integrates insights from organizational behavior, artificial intelligence, and socio-technical systems theory to develop a unified framework, thereby bridging previously disconnected streams of literature.

By shifting the analytical focus from adoption to alignment, this paper offers a new perspective on the challenges of AI integration. As organizations increasingly rely on intelligent systems, the critical question is no longer whether AI should be adopted, but how human and machine systems can be effectively aligned. Without such alignment, the potential benefits of AI may remain unrealized, or even lead to unintended negative consequences for individuals and organizations alike.

2. Theoretical Foundations

2.1 AI in Organizations: From Automation to Augmentation

The role of artificial intelligence (AI) in organizations has evolved significantly, shifting from a narrow focus on automation toward broader forms of augmentation. Early research conceptualized AI primarily as a substitute for human labor, emphasizing its ability to automate routine and predictable tasks, thereby improving efficiency and reducing costs (Autor, 2015; Frey & Osborne, 2017; Acemoglu & Restrepo, 2020). This perspective aligns with economic models that frame AI as a labor-displacing technology, particularly in structured and rule-based environments (Agrawal et al., 2019; Bessen, 2019).

More recent scholarship, however, challenges this substitution logic by highlighting the complementary role of AI in enhancing human capabilities. Rather than replacing workers, AI increasingly functions as a cognitive partner, enabling new forms of problem-solving, decision-making, and knowledge generation (Raisch & Krakowski, 2021; Brynjolfsson et al., 2021). This transition has given rise to the concept of hybrid intelligence, where human and machine capabilities are combined to achieve superior outcomes compared to either acting alone (Dellermann et al., 2019; Wilson & Daugherty, 2018). In this context, AI is no longer a passive tool but an active participant in organizational processes, influencing how decisions are made and how expertise is defined (Faraj et al., 2018; Berente et al., 2021).

Importantly, this shift toward augmentation introduces new complexities in organizational design. AI-driven systems alter task interdependence, redistribute decision authority, and redefine performance evaluation mechanisms (Tambe et al., 2019; Davenport et al., 2020). These changes are not linear; instead, organizations often experience transitional inefficiencies, as reflected in the "productivity J-curve," where the benefits of AI adoption emerge only after significant adaptation and learning (Brynjolfsson et al., 2021). As such, understanding AI in organizations requires moving beyond static views of technology toward dynamic perspectives that account for continuous adaptation and co-evolution between humans and intelligent systems (Keding, 2021; von Krogh, 2018).

2.2 Human–AI Interaction as a Socio-Technical System

The interaction between humans and AI is best understood through the lens of socio-technical systems theory, which emphasizes the interdependence between social and technological components in shaping organizational outcomes (Cascio & Montealegre, 2016; Parker et al., 2017). Within this perspective, AI is not treated as an external tool but as an embedded actor that co-evolves with human practices, organizational routines, and institutional structures (Faraj et al., 2018; Murray et al., 2021).

Recent research extends this view by conceptualizing human–AI interaction as a form of conjoined agency, where decision-making authority is distributed across human and machine actors (Murray et al., 2021; Raisch & Fomina, 2024). This has led to the emergence of concepts such as human–AI teaming and hybrid collaboration, where AI systems contribute analytical and predictive capabilities, while humans provide contextual understanding, ethical reasoning, and social interpretation (Seeber et al., 2020; Jarrahi, 2018). The effectiveness of such collaboration depends on the extent to which these complementary capabilities are aligned and mutually reinforcing (Fügener et al., 2022).

However, achieving alignment within socio-technical systems is inherently challenging. AI systems often operate based on probabilistic models and opaque algorithms, which may not align with human expectations of transparency and interpretability (Burrell, 2016; Arrieta et al., 2020). This opacity creates coordination challenges, as users struggle to understand, evaluate, and justify AI-generated outputs (Lebovitz et al., 2022). Furthermore, the delegation of decision-making authority to AI introduces ambiguity in responsibility attribution, complicating accountability structures within organizations (Shrestha et al., 2019; Kim & Hinds, 2006).

These dynamics suggest that human–AI systems should be conceptualized as dynamic and evolving assemblages, where both humans and AI continuously adapt to each other. Failures in this co-adaptation process often result in coordination breakdowns, even when individual components perform effectively (Amershi et al., 2019; de Vreede et al., 2021). This highlights the importance of examining not only technological capabilities but also the relational processes that govern human–AI interaction.

2.3 Behavioral Responses to AI: Algorithm Aversion and Appreciation

A substantial body of research has examined how individuals respond behaviorally to AI systems, particularly in decision-making contexts. One of the most widely documented phenomena is algorithm aversion, where individuals prefer human judgment over algorithmic recommendations, even when the latter demonstrate superior performance (Dietvorst et al., 2015; Prahl & Van Swol, 2017). This aversion is often driven by perceptions that algorithms lack flexibility, contextual sensitivity, or the ability to account for unique circumstances (Castelo et al., 2019).

In contrast, research on algorithm appreciation shows that individuals may prefer AI over human judgment under certain conditions, particularly when tasks are perceived as objective, data-intensive, and less dependent on subjective interpretation (Logg et al., 2019; Yeomans et al., 2019). These seemingly contradictory findings indicate that human responses to AI are not fixed but context-dependent, shaped by factors such as task characteristics, perceived competence, and the degree of user control (Burton et al., 2020).

Further research suggests that individuals are more willing to accept AI recommendations when they are allowed to interact with or modify system outputs (Dietvorst et al., 2018). This highlights the importance of perceived control in shaping user behavior. However, when AI systems operate as "black boxes," users often experience uncertainty and discomfort, leading to resistance or disengagement (Kizilcec, 2016; Liao et al., 2020). Taken together, these findings suggest that behavioral responses to AI are dynamic and contingent, emerging from the interaction between system design and human cognition (Araujo et al., 2020).

2.4 Trust, Opacity, and Explainability in AI Systems

Trust plays a central role in shaping human–AI interaction. It determines whether individuals accept, rely on, or reject AI-generated recommendations (Hoff & Bashir, 2015; Glikson & Woolley, 2020). However, trust in AI differs from traditional forms of interpersonal trust, as it is influenced by both technological attributes and psychological perceptions (Cabiddu et al., 2022).

One of the primary challenges in building trust is algorithmic opacity. Many AI systems rely on complex machine learning models that are difficult to interpret, creating a gap between system functionality and user understanding (Burrell, 2016). This opacity can lead to perceptions of unfairness and loss of control, particularly in high-stakes decision-making contexts (Binns et al., 2018; Lee, 2018). As a result, users may hesitate to rely on AI, even when it produces accurate outcomes (Lebovitz et al., 2022).

Efforts to address these challenges have led to the development of explainable AI (XAI), which aims to make AI systems more transparent and interpretable (Arrieta et al., 2020; Haque et al., 2023). While XAI improves user understanding, it does not fully resolve the tension between model complexity and human cognitive limitations. In some cases, increased transparency can even overwhelm users, reducing confidence and trust (Endsley, 2017; Liao et al., 2020). These findings suggest that trust in AI is a dynamic and context-dependent construct, shaped by the interplay between system design, user cognition, and situational factors.

2.5 Algorithmic Management and the Transformation of Work

The integration of AI into organizational systems has also given rise to algorithmic management, where decision-making and control are increasingly mediated by data-driven systems (Kellogg et al., 2020; Lee et al., 2015). In such environments, employees are monitored, evaluated, and directed through algorithms, often with limited human oversight (Gray & Suri, 2019).

While algorithmic management can enhance efficiency and consistency, it also introduces significant challenges related to autonomy, fairness, and identity (Wright & Schultz, 2018; Grote & Berens, 2020). Employees may perceive themselves as reduced to data points, leading to diminished motivation and engagement (Brougham & Haar, 2018). Additionally, the absence of human interaction in decision-making processes can create feelings of alienation and psychological detachment (Puntoni et al., 2021).

These dynamics are particularly problematic in roles where identity and meaning are central to performance. When AI systems dominate decision-making processes, individuals may struggle to find purpose and significance in their work (Bankins & Formosa, 2023). This highlights a critical limitation of current AI-driven work systems: their inability to account for the human need for meaning, autonomy, and social recognition.

2.6 Synthesis and Research Gap

Taken together, the existing literature provides a rich but fragmented understanding of human–AI interaction. Research on AI capabilities emphasizes technological performance and organizational transformation (Davenport et al., 2020; Berente et al., 2021), while behavioral studies focus on individual responses such as trust and resistance (Dietvorst et al., 2015; Logg et al., 2019). Meanwhile, socio-technical perspectives highlight the importance of alignment between human and technological systems (Cascio & Montealegre, 2016; Seeber et al., 2020).

A structured clarification of constructs is necessary to prevent ambiguity and ensure that each element of the framework is analytically distinct and theoretically grounded. The following table consolidates definitions, theoretical roots, and roles within the model, thereby stabilizing the conceptual architecture before the framework is operationalized.

Table 1. Construct Definitions and Theoretical Anchors in the Human–AI Misalignment Framework

Construct	Definition	Theoretical Foundation	Role in Model
AI System Characteristics	Technical attributes of AI systems, including opacity, autonomy, and data-driven decision logic	Artificial intelligence research; algorithmic decision-making literature	Antecedent
Work Context Conditions	Task and organizational features shaping AI use, such as complexity and interdependence	Socio-technical systems theory; organizational design literature	Antecedent
Human–AI Misalignment	Systematic discrepancy between AI operational logic and human cognitive, emotional, and social structures	Socio-technical systems theory; hybrid intelligence literature	Core Construct
Cognitive Misalignment	Incompatibility between statistical inference and human interpretive reasoning	Cognitive psychology; bounded rationality theory	Dimension of Misalignment
Emotional Misalignment	Lack of alignment between AI outputs and human affective and empathetic expectations	Affective science; human–computer interaction literature	Dimension of Misalignment
Agency Misalignment	Discrepancy in perceived control and decision	Agency theory; human–automation interaction	Dimension of Misalignment

Construct	Definition	Theoretical Foundation	Role in Model
	authority between humans and AI systems		
Meaning Misalignment	Divergence between AI-driven efficiency goals and human need for purpose and identity	Meaningful work literature; organizational behavior	Dimension of Misalignment
Cognitive Overload	Excessive mental processing demands caused by complex or opaque AI outputs	Cognitive load theory; human factors research	Psychological Mechanism
Identity Threat	Perceived challenge to professional identity due to AI involvement in decision-making	Identity theory; organizational behavior	Psychological Mechanism
Perceived Loss of Control	Reduced sense of autonomy in decision processes involving AI	Self-determination theory; automation research	Psychological Mechanism
Perceived Unfairness	Perception that AI decisions lack transparency or justice	Organizational justice theory; algorithmic fairness literature	Psychological Mechanism
Algorithm Aversion	Tendency to reject AI recommendations despite superior performance	Behavioral decision-making literature	Behavioral Outcome
Blind Reliance	Uncritical acceptance of AI outputs without evaluation	Automation bias literature	Behavioral Outcome
Resistance	Active opposition to AI adoption or use	Change management; technology acceptance research	Behavioral Outcome
Disengagement	Withdrawal from interaction with AI systems	Work motivation; organizational behavior	Behavioral Outcome
Organizational Performance	Effectiveness of outcomes resulting from human–AI collaboration	Strategic management; performance literature	Outcome
Trust in AI	Willingness to rely on AI systems based on perceived reliability and competence	Trust in automation literature	Moderator
Psychological Safety	Perceived safety in experimenting with AI without negative consequences	Organizational behavior; team learning theory	Moderator
AI Literacy	Individual capability to understand and interact with AI systems	Digital literacy; AI adoption research	Moderator
Leadership Support	Organizational guidance and endorsement of AI use	Leadership theory; change management	Moderator

Source: Developed by the authors

By consolidating dispersed constructs into a single structured reference, Table 1 enhances conceptual precision and reduces the risk of overlap or implicit redefinition across sections. Table 1 also clarifies the analytical role of each construct within the framework, enabling readers to distinguish clearly between antecedents, core constructs, mechanisms, outcomes, and moderators, thereby strengthening the internal coherence of the overall model.

Despite these contributions, a critical gap remains. Existing research has not fully theorized the mechanisms through which AI system characteristics translate into psychological processes and behavioral outcomes. As a result, the field lacks an integrated explanation of why human–AI collaboration often fails, even in technologically advanced settings (Berente et al., 2021; Lebovitz et al., 2022).

To address this limitation, this study develops a mechanism-based conceptual framework of Human–AI Misalignment. By integrating insights from artificial intelligence, organizational behavior, and socio-technical systems theory, the framework provides a systematic explanation of how misalignment emerges and how it shapes human behavior and organizational outcomes.

3. Conceptual Framework: Human–AI Misalignment

3.1 Defining Human–AI Misalignment

The growing integration of artificial intelligence into organizational systems has created new forms of interaction between humans and intelligent technologies. While prior research highlights the performance benefits of AI-enabled decision-making and hybrid intelligence systems, empirical evidence consistently shows that these benefits are not uniformly realized in practice (Raisch & Krakowski, 2021; Davenport et al., 2020; Brynjolfsson et al., 2021). Instead, organizations frequently encounter friction, inconsistency, and breakdowns in collaboration when humans interact with AI systems (Dietvorst et al., 2015; Logg et al., 2019; Longoni et al., 2019). These observations suggest that existing perspectives centered on adoption and capability are insufficient to explain the underlying dynamics of human–AI interaction.

To address this limitation, this study introduces the construct of **Human–AI Misalignment**, defined as a systematic discrepancy between the operational logic of AI systems and the cognitive, emotional, and social structures of human actors, resulting in dysfunctional collaboration outcomes within organizational contexts. This definition builds on emerging perspectives that conceptualize AI as an embedded socio-technical actor rather than a neutral tool (Faraj et al., 2018; Berente et al., 2021; Murray et al., 2021). Within this view, both humans and AI contribute to decision-making processes, yet operate according to fundamentally different principles.

AI systems rely on probabilistic modeling, pattern recognition, and large-scale data processing to generate predictions and optimize outcomes (Agrawal et al., 2019; Arrieta et al., 2020). In contrast, human cognition is shaped by bounded rationality, heuristics, and contextual interpretation, often prioritizing interpretability and narrative coherence over statistical optimality (Jarrahi, 2018; Castelo et al., 2019). This divergence reflects an epistemic gap between statistical inference and human sensemaking, which can lead to interpretive tension and reduced alignment in decision processes (Burrell, 2016; Yeomans et al., 2019).

Importantly, Human–AI Misalignment is not conceptualized as a binary condition, but as a multidimensional and dynamic phenomenon that evolves through interaction. Misalignment emerges when discrepancies between AI outputs and human expectations are not effectively reconciled, leading to breakdowns in trust, coordination, and performance (Glikson & Woolley, 2020; Lebovitz et al., 2022). By framing misalignment as a systemic issue, this construct shifts the analytical focus from individual-level biases toward relational processes within socio-technical systems (Cascio & Montealegre, 2016; Seeber et al., 2020).

Based on the conceptualization developed above, this study proposes a mechanism-based framework of Human–AI Misalignment. The framework integrates insights from artificial intelligence, organizational behavior, and socio-technical systems theory to explain how discrepancies between AI system logic and human cognitive, emotional, and social structures generate misalignment. This misalignment subsequently activates a set of psychological mechanisms that shape behavioral responses and ultimately influence organizational performance.

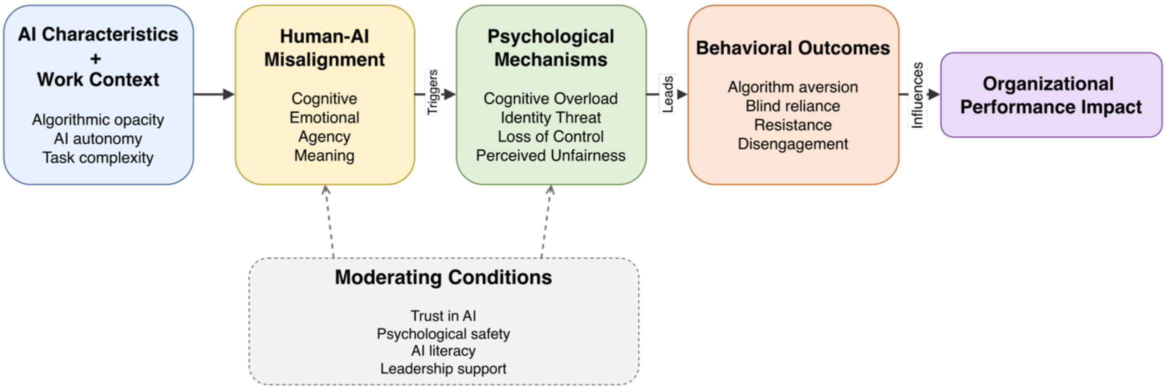

Figure 1. A Mechanism-Based Conceptual Framework of Human–AI Misalignment
Source: Developed by the authors

As illustrated in Figure 1, the framework begins with AI system characteristics and work context conditions, which shape the emergence of Human–AI Misalignment across cognitive, emotional, agency-related, and meaning-based dimensions. These dimensions trigger psychological mechanisms, including cognitive overload, identity threat, perceived loss of control, and perceived unfairness. These mechanisms lead to a range of behavioral outcomes, such as algorithm aversion, blind reliance, resistance, and disengagement, which in turn impact organizational performance. The model also highlights the role of moderating conditions, including trust in AI, psychological safety, AI literacy, and leadership support, which influence the strength of the relationships between misalignment, psychological mechanisms, and behavioral outcomes.

3.2 Dimensions of Human–AI Misalignment

Human–AI Misalignment manifests across four interrelated dimensions that reflect fundamental differences in cognition, emotion, agency, and meaning between humans and AI systems.

A structured representation is necessary to isolate the internal architecture of Human–AI Misalignment and prevent it from being treated as a monolithic construct. The following figure analytically decomposes misalignment into four parallel dimensions, each rooted in a distinct form of discrepancy between human and AI systems.

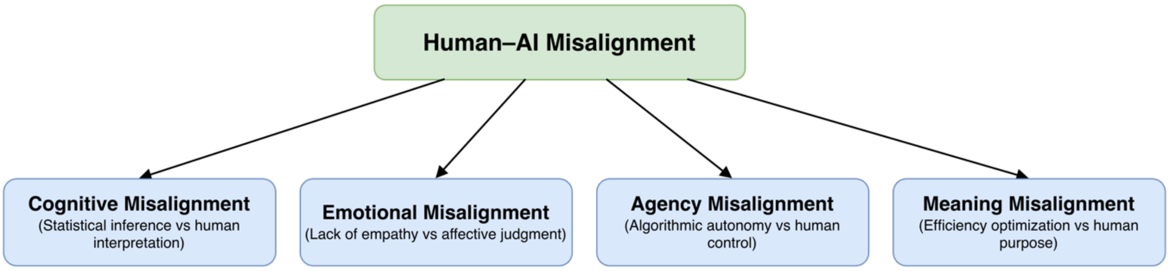

Figure 2. Dimensions of Human–AI Misalignmen
Source: Developed by the authors

The structure articulated in Figure 2 clarifies that Human–AI Misalignment is not a singular construct but a configuration of four analytically distinct yet coexisting discrepancies. Each dimension captures a specific incompatibility between AI system logic and human cognitive, emotional, agentic, and meaning-making processes.

a. Cognitive misalignment

Cognitive misalignment arises from differences in how humans and AI process information and generate decisions. AI systems produce outputs based on statistical inference and optimization, often achieving high levels of predictive accuracy (Arrieta et al., 2020; Brynjolfsson et al., 2021). However, these outputs are frequently difficult for users to interpret, particularly when based on complex and opaque models (Burrell, 2016; Kizilcec, 2016). In contrast, humans rely on heuristics and mental models that prioritize simplicity, coherence, and contextual relevance (Kahneman, 2011; Castelo et al., 2019).

This discrepancy creates challenges in understanding and evaluating AI recommendations. When users cannot interpret the rationale behind algorithmic outputs, they may experience cognitive overload and uncertainty, leading to rejection or misuse of AI systems (Liao et al., 2020; Burton et al., 2020). Empirical studies show that even minor algorithmic errors can disproportionately reduce trust, despite overall superior performance (Dietvorst et al., 2015). This suggests that cognitive misalignment is rooted not only in knowledge gaps but in fundamental incompatibilities between human interpretive frameworks and machine logic.

b. Emotional misalignment

Human–AI interaction is also shaped by emotional processes, which are largely absent in AI systems. While AI can simulate certain affective cues, it lacks genuine emotional understanding and empathy, which play a critical role in human decision-making (Huang & Rust, 2018; Puntoni et al., 2021). This creates emotional misalignment, where users perceive AI systems as impersonal, unempathetic, or socially disconnected.

Such misalignment is particularly salient in contexts involving moral judgment, interpersonal interaction, or high personal relevance. Individuals are less willing to accept AI decisions in these domains, even when they are objectively accurate (Longoni et al., 2019; Granulo et al., 2019). Emotional responses such as anxiety, discomfort, and distrust further reinforce resistance to AI adoption (Araujo et al., 2020; Brougham & Haar, 2018). These findings highlight the importance of affective alignment in shaping the effectiveness of human–AI collaboration.

c. Agency misalignment

Agency misalignment reflects discrepancies in perceived control, responsibility, and decision authority between humans and AI systems. As AI becomes more autonomous, decision-making authority is increasingly distributed across human and machine actors, creating ambiguity in accountability structures (Shrestha et al., 2019; Murray et al., 2021). This challenges traditional notions of agency, where humans are assumed to be the primary decision-makers.

In some cases, individuals may over-rely on AI, delegating decisions without sufficient critical evaluation, a phenomenon often associated with automation bias (Endsley, 2017). In other cases, users may resist AI recommendations due to perceived loss of control or authority (Prahl & Van Swol, 2017). Both patterns reflect difficulties in calibrating appropriate reliance on AI systems. Algorithmic management further intensifies this issue by shifting control from human supervisors to data-driven systems, often reducing autonomy and increasing surveillance (Kellogg et al., 2020; Lee et al., 2015). These dynamics suggest that agency in human–AI systems is negotiated rather than fixed, and misalignment arises when this negotiation fails.

d. Meaning misalignment

Meaning misalignment captures differences in how humans and AI systems interpret the purpose and significance of work. AI systems are designed to optimize efficiency, accuracy, and performance, whereas humans seek meaning, identity, and intrinsic motivation in their work (Bankins & Formosa, 2023; Puntoni et al., 2021). This divergence can create a sense of alienation, particularly in environments dominated by algorithmic decision-making.

Employees may perceive their roles as reduced to executing tasks defined by AI systems, undermining their sense of autonomy and contribution (Gray & Suri, 2019). Over time, this can erode intrinsic motivation and engagement, particularly in roles with high identity salience (Brougham & Haar, 2018). Meaning misalignment thus highlights a critical limitation of AI-driven work systems, namely their inability to account for the existential and identity-related dimensions of human work.

3.3 Mechanisms Linking Misalignment to Behavioral Outcomes

To explain how Human–AI Misalignment translates into observable behaviors, this study proposes a set of psychological mechanisms that mediate this relationship. These mechanisms function as cognitive and affective processes that transform structural discrepancies into behavioral responses (Berente et al., 2021; Lebovitz et al., 2022).

A more granular articulation of the causal logic is required to clarify how misalignment is translated into observable behavior through specific psychological processes. The following figure isolates this transformation layer by mapping each dimension of misalignment to its corresponding mechanism, thereby making the internal causal pathways explicit rather than implicit.

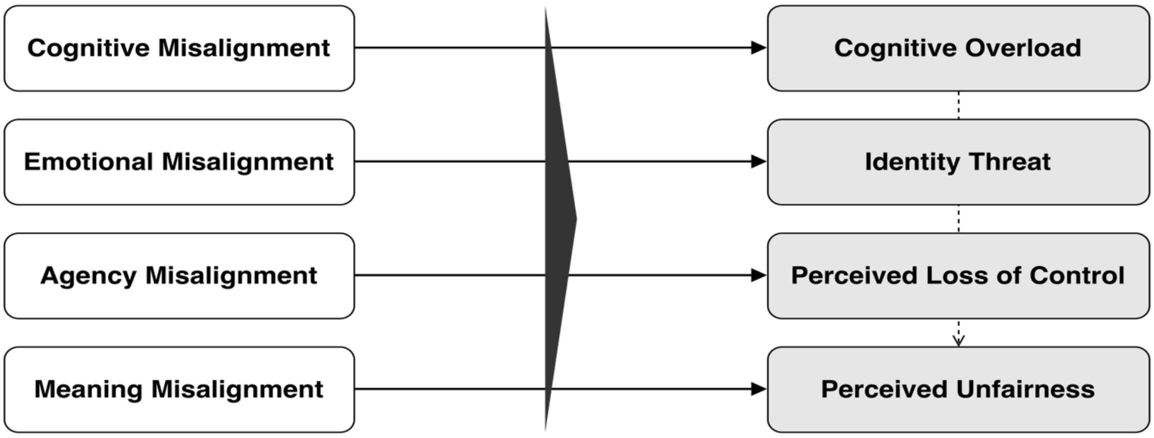

Figure 3. Psychological Mechanisms Linking Human–AI Misalignment to Behavioral Responses
Source: Developed by the author

Figure 3 isolates the mechanism layer by specifying how each dimension of Human–AI Misalignment activates a corresponding psychological process. Rather than assuming a direct relationship between misalignment and behavior, the figure clarifies that cognitive discrepancies produce cognitive overload, emotional discrepancies trigger identity threat, agency-related discrepancies generate perceived loss of control, and meaning-based discrepancies lead to perceived unfairness. By making these pathways explicit, Figure 3 strengthens the causal logic of the model and provides a clearer basis for explaining variability in behavioral responses across different human–AI interaction contexts.

Cognitive misalignment is expected to increase cognitive overload, as individuals struggle to interpret complex and opaque AI outputs (Endsley, 2017; Liao et al., 2020). Emotional misalignment contributes to identity threat, particularly in professional contexts where expertise and judgment are central to self-concept (Lebovitz et al., 2022; Puntoni et al., 2021). Agency misalignment leads to perceived loss of control, as individuals experience reduced autonomy in decision-making processes (Shrestha et al., 2019; Murray et al., 2021). Finally, meaning misalignment increases perceptions of unfairness and alienation, particularly when AI systems produce outcomes that lack transparency or social legitimacy (Binns et al., 2018; Lee, 2018).

These mechanisms are not independent but interact in complex ways, shaping how individuals respond to AI systems. For example, cognitive overload may amplify perceptions of unfairness, while identity threat may intensify resistance behaviors. Understanding these mechanisms is therefore critical for explaining the variability and inconsistency of human responses to AI.

3.4 Behavioral Outcomes of Human–AI Misalignment

The psychological mechanisms described above give rise to a range of behavioral outcomes that affect the effectiveness of human–AI collaboration. One key outcome is algorithm aversion, where individuals reject AI recommendations despite their accuracy (Dietvorst et al., 2015; Castelo et al., 2019). Conversely, some users exhibit blind reliance, accepting AI outputs without critical evaluation, particularly under conditions of high cognitive load (Endsley, 2017; Yeomans et al., 2019).

A more fine-grained differentiation of behavioral outcomes is necessary to avoid treating responses to AI as uniform or interchangeable. The following table analytically maps each behavioral outcome to its dominant psychological mechanism and underlying form of misalignment, thereby clarifying why different patterns of response emerge under different conditions.

Table 2. Behavioral Outcomes of Human–AI Misalignment and Their Underlying Mechanisms

Behavioral Outcome	Dominant Psychological Mechanism	Underlying Misalignment Dimension	Typical Organizational Effect
Algorithm Aversion	Cognitive Overload	Cognitive Misalignment	Underutilization of AI capabilities despite high accuracy
Blind Reliance	Cognitive Overload; Perceived Loss of Control	Agency Misalignment	Overdependence on AI, reduced critical evaluation
Resistance	Identity Threat	Emotional Misalignment	Barriers to adoption and implementation failure
Disengagement	Perceived Unfairness	Meaning Misalignment	Reduced engagement, withdrawal from AI interaction
Selective Use	Mixed mechanisms depending on context	Multiple misalignment dimensions	Inconsistent AI usage and unstable performance outcomes
Adaptive Collaboration	Managed cognitive load; reduced identity threat	Low or mitigated misalignment	Effective human–AI integration and improved decision quality

Source: Developed by the authors

By explicitly linking behavioral outcomes to their underlying mechanisms and misalignment dimensions, Table 2 clarifies that responses to AI are not random but systematically generated through distinct psychological pathways. Table 2 strengthens the explanatory power of the framework by demonstrating how different configurations of misalignment produce qualitatively different behavioral patterns, thereby enabling more precise interpretation of human–AI interaction outcomes.

Other outcomes include resistance to AI adoption, where individuals actively oppose the use of AI systems, and disengagement, where users withdraw from interaction with AI altogether (Brougham & Haar, 2018; Puntoni et al., 2021). These behaviors reduce the effectiveness of human–AI collaboration and may lead to suboptimal decision-making and performance outcomes.

Importantly, these behavioral responses are not mutually exclusive and may coexist within the same organizational context. For instance, individuals may exhibit both aversion and over-reliance depending on situational factors, highlighting the dynamic nature of human–AI interaction (Burton et al., 2020; Araujo et al., 2020).

3.5 Moderating Conditions

The impact of Human–AI Misalignment on behavioral outcomes is influenced by several moderating factors that shape how individuals interpret and respond to AI systems. Trust in AI plays a central role, as higher levels of trust reduce resistance and increase willingness to rely on AI outputs (Glikson & Woolley, 2020; Hoff & Bashir, 2015). Psychological safety enables individuals to experiment with AI systems without fear of negative consequences, thereby reducing identity threat and resistance (Edmondson, 1999).

AI literacy also plays a critical role by enhancing users' ability to understand and interpret AI outputs, thereby mitigating cognitive misalignment (Haque et al., 2023). Leadership support further influences the adoption process by providing guidance, legitimacy, and resources for effective human–AI collaboration (Tambe et al., 2019; Fountaine et al., 2019).

These moderating conditions highlight the importance of organizational context in shaping the outcomes of human–AI interaction. Misalignment is not solely determined by technological characteristics but is influenced by social, cultural, and institutional factors that affect how AI is perceived and used.

4. Discussion

4.1 From Adoption to Alignment

This study challenges the dominant assumption that successful AI integration is primarily determined by adoption, usability, or technological sophistication. While prior research emphasizes system performance and user acceptance as key drivers of AI success, the findings presented here suggest that these perspectives are insufficient to explain persistent dysfunctions in human–AI collaboration (Davenport et al., 2020; Tambe et al., 2019; Berente et al., 2021). Instead, the central issue lies not in whether AI is adopted, but in whether it is aligned with human cognitive, emotional, and social structures.

By introducing the concept of Human–AI Misalignment, this study shifts the analytical focus from adoption outcomes to interaction processes within socio-technical systems. This reframing highlights that even highly advanced AI systems may fail to generate value when their operational logic is incompatible with human interpretive frameworks (Raisch & Krakowski, 2021; Murray et al., 2021). The findings thus extend existing work on the automation and augmentation paradox by demonstrating that augmentation is not inherently beneficial, but contingent on the degree of alignment between human and machine systems (Brynjolfsson et al., 2021; Dellermann et al., 2019).

This perspective also contributes to a more nuanced understanding of AI-enabled work systems. Rather than viewing performance outcomes as a function of technological capability alone, this study positions them as emergent properties of relational dynamics between humans and AI. In doing so, it responds to recent calls for process-oriented theorizing in AI research by providing a mechanism-based explanation of collaboration breakdowns (Lebovitz et al., 2022; Fügener et al., 2022).

4.2 Theoretical Contributions

a. Reconceptualizing socio-technical systems

This study advances socio-technical systems theory by reconceptualizing the role of technology within organizational contexts. Traditional perspectives emphasize the need for alignment between social and technical subsystems, but often treat technology as relatively stable and predictable (Cascio & Montealegre, 2016; Parker et al., 2017). In contrast, this study positions AI as a dynamic and semi-autonomous actor that actively shapes human cognition, decision-making, and organizational processes (Faraj et al., 2018; Murray et al., 2021).

By introducing Human–AI Misalignment as a multidimensional construct, this research extends socio-technical theory beyond alignment as a static goal toward alignment as an ongoing and negotiated process. This perspective captures the inherent tensions that arise when human and machine logics interact, providing a more realistic account of contemporary organizational systems characterized by hybrid intelligence (Seeber et al., 2020; Raisch & Fomina, 2024).

b. Bridging organizational behavior and AI research

A second contribution lies in bridging the gap between organizational behavior and AI research. Existing studies in organizational behavior have extensively examined constructs such as trust, identity, and motivation, yet these constructs are rarely integrated into AI-focused research (Glikson & Woolley, 2020; Puntoni et al., 2021). Conversely, AI research often focuses on system performance and adoption, overlooking the psychological processes that shape human responses (Berente et al., 2021; Davenport et al., 2020).

This study integrates these perspectives by introducing a mechanism-based model that links AI characteristics to psychological processes and behavioral outcomes. Specifically, it identifies cognitive overload, identity threat, perceived loss of control, and perceived unfairness as key mediating mechanisms. This integration contributes to theory development by moving beyond descriptive accounts toward causal explanations of human–AI interaction (Lebovitz et al., 2022; Burton et al., 2020).

c. Introducing Human–AI Misalignment as a core construct

Third, this study introduces Human–AI Misalignment as a theoretically grounded construct that captures the systemic and multidimensional nature of breakdowns in human–AI collaboration. Unlike existing concepts such as algorithm aversion or trust in AI, which focus on specific aspects of human response, misalignment provides an integrative lens that connects cognitive, emotional, agency-related, and meaning-based dimensions (Dietvorst et al., 2015; Hoff & Bashir, 2015).

By conceptualizing misalignment as a mediator between AI system characteristics and behavioral outcomes, this study offers a new framework for analyzing the impact of AI on organizational performance. This opens up opportunities for future research, including empirical testing, scale development, and cross-context analysis. The construct thus provides both theoretical depth and practical applicability, enhancing its relevance for ongoing debates in AI and organizational research.

4.3 Managerial Implications

a. Designing for alignment rather than efficiency

The findings of this study suggest that organizations must move beyond efficiency-driven approaches to AI implementation and adopt a more human-centered perspective. While AI systems are often designed to optimize accuracy and scalability, insufficient attention is given to how these systems align with human cognition and behavior (Davenport et al., 2020; Arrieta et al., 2020).

Organizations should therefore prioritize design principles that enhance interpretability, provide meaningful feedback, and enable user control. Such features can reduce cognitive and agency misalignment, thereby improving the effectiveness of human–AI collaboration (Liao et al., 2020; Fügener et al., 2022).

b. Building trust and psychological safety

Trust emerges as a critical factor in mitigating the negative effects of misalignment. Organizations must actively cultivate trust in AI systems by ensuring transparency, fairness, and reliability (Glikson & Woolley, 2020; Cabiddu et al., 2022). At the same time, fostering psychological safety enables employees to experiment with AI systems, learn from errors, and develop confidence in their interactions (Edmondson, 1999).

c. Developing AI literacy and adaptive capabilities

Another key implication is the need to invest in AI literacy and adaptive capabilities. Employees must be equipped not only with technical skills but also with the ability to interpret, question, and collaborate with AI systems (Haque et al., 2023). This requires continuous learning approaches that integrate technical training with cognitive and behavioral development.

d. Redesigning work and preserving meaning

Finally, organizations must address meaning misalignment by redesigning work in ways that preserve human agency, identity, and purpose. AI should be integrated in ways that complement human strengths rather than replace them, ensuring that employees continue to experience meaningful engagement in their roles (Bankins & Formosa, 2023; Gray & Suri, 2019).

4.4 Policy and Societal Implications

The implications of Human–AI Misalignment extend beyond organizational boundaries, raising important questions for policymakers and society. As AI becomes increasingly embedded in decision-making processes, issues related to transparency, fairness, and accountability become more salient (Binns et al., 2018; Wright & Schultz, 2018).

Policymakers must develop regulatory frameworks that balance innovation with ethical considerations. This includes establishing standards for explainability, ensuring accountability in algorithmic decisions, and protecting individuals from potential harms associated with misalignment (Grote & Berens, 2020). Addressing these challenges is essential for building public trust in AI and ensuring its responsible deployment.

4.5 Future Research Agenda

This study opens several avenues for future research. First, empirical studies are needed to validate the proposed framework and test the relationships between misalignment dimensions, psychological mechanisms, and behavioral outcomes. This includes the development of measurement scales for Human–AI Misalignment, enabling quantitative analysis across different contexts.

Second, future research should examine how misalignment evolves over time, particularly as individuals and organizations gain experience with AI systems. Longitudinal studies can provide insights into the dynamics of alignment and adaptation.

Third, cross-context and cross-cultural studies are needed to explore how differences in institutional environments, organizational cultures, and societal norms influence human–AI interaction. Such research can enhance the generalizability of the framework and identify context-specific factors that shape alignment.

5. Conclusion

The rapid advancement of artificial intelligence has fundamentally transformed how organizations design work, make decisions, and create value. Yet, as this study demonstrates, the central challenge is no longer whether organizations can adopt AI, but whether they can effectively align human and machine systems within increasingly complex socio-technical environments.

By introducing the concept of Human–AI Misalignment, this paper offers a new lens for understanding why human–AI collaboration often fails despite significant technological progress. Rather than interpreting resistance, distrust, or misuse of AI as isolated behavioral anomalies, this study conceptualizes these outcomes as manifestations of deeper misalignments across cognitive, emotional, agency-related, and meaning-based dimensions. Through a mechanism-based framework, it explains how these misalignments emerge from the interaction between AI system characteristics and work context conditions, and how they translate into psychological processes and behavioral outcomes that ultimately shape organizational performance.

This perspective contributes to the literature by shifting the focus from adoption to alignment, advancing socio-technical systems theory, and bridging organizational behavior with artificial intelligence research. It also provides a foundation for future empirical work by offering a structured framework that can be tested, refined, and extended across different organizational and cultural contexts.

Beyond its theoretical contributions, this study carries a broader implication. The success of artificial intelligence in organizations will not be determined solely by the sophistication of algorithms, but by the extent to which these systems are designed and managed in harmony with human cognition, emotion, and social meaning. In this sense, alignment is not a secondary consideration, but the defining condition of effective human–AI collaboration.

As AI continues to evolve, the future of work will depend not only on technological innovation, but on our ability to rethink and redesign the relationship between humans and intelligent systems.

References

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Artificial intelligence: The ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives, 33*(2), 31–50. https://doi.org/10.1257/jep.33.2.31

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society, 35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Autor, D. H. (2015). Why are there still so many jobs? *Journal of Economic Perspectives, 29*(3), 3–30. https://doi.org/10.1257/jep.29.3.3

Bankins, S., & Formosa, P. (2023). The ethical implications of artificial intelligence for meaningful work. *Journal of Business Ethics, 185*, 725–740. https://doi.org/10.1007/s10551-022-05097-8

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly, 45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Bessen, J. E. (2019). AI and jobs. *NBER Working Paper*. https://doi.org/10.3386/w24235

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). Perceptions of justice in algorithmic decisions. *CHI Proceedings*, 1–14. https://doi.org/10.1145/3173574.3173951

Brougham, D., & Haar, J. (2018). Smart technology and employee perceptions. *Journal of Management & Organization, 24*(2), 239–257. https://doi.org/10.1017/jmo.2016.55

Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve. *American Economic Journal: Macroeconomics, 13*(1), 333–372. https://doi.org/10.1257/mac.20180386

Burrell, J. (2016). How the machine "thinks". *Big Data & Society, 3*(1). https://doi.org/10.1177/2053951715622512

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). Algorithm aversion review. *Journal of Behavioral Decision Making, 33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Cabiddu, F., Moi, L., & Patriotta, G. (2022). Trust in algorithms. *European Management Journal, 40*(5), 685–706. https://doi.org/10.1016/j.emj.2022.06.001

Cascio, W. F., & Montealegre, R. (2016). Technology and organizations. *Annual Review of Organizational Psychology, 3*, 349–375. https://doi.org/10.1146/annurev-orgpsych-041015-062352

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Algorithm aversion. *Journal of Marketing Research, 56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). AI in marketing. *Journal of the Academy of Marketing Science, 48*, 24–42. https://doi.org/10.1007/s11747-019-00696-0

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering, 61*, 637–643. https://doi.org/10.1007/s12599-019-00595-2

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion. *Journal of Experimental Psychology, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Edmondson, A. C. (1999). Psychological safety. *Administrative Science Quarterly, 44*(2), 350–383. https://doi.org/10.2307/2666999

Endsley, M. R. (2017). Human–automation interaction. *Human Factors, 59*(1), 5–27. https://doi.org/10.1177/0018720816681350

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Learning algorithms. *Information and Organization, 28*(1), 62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

Fountaine, T., McCarthy, B., & Saleh, T. (2019). AI-powered organizations. *Harvard Business Review, 97*(4), 62–73

Frey, C. B., & Osborne, M. A. (2017). Future of employment. *Technological Forecasting, 114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Human-AI collaboration. *Information Systems Research, 33*(2), 678–696. https://doi.org/10.1287/isre.2021.1079

Glikson, E., & Woolley, A. W. (2020). Trust in AI. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Gray, M. L., & Suri, S. (2019). *Ghost work*. Houghton Mifflin

Grote, G., & Berens, P. (2020). Ethics in AI. *Journal of Medical Ethics, 46*(3), 205–211. https://doi.org/10.1136/medethics-2019-105586

Haque, A. K. M. B., Islam, A. K. M. N., Mikalef, P., & Tarafdar, M. (2023). Explainable AI. *Technological Forecasting, 186*, 122120. https://doi.org/10.1016/j.techfore.2022.122120

Hoff, K. A., & Bashir, M. (2015). Trust in automation. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Huang, M.-H., & Rust, R. T. (2018). AI in service. *Journal of Service Research, 21*(2), 155–172. https://doi.org/10.1177/1094670517752459

Huang, M.-H., & Rust, R. T. (2021). AI strategy. *Journal of the Academy of Marketing Science, 49*, 30–50. https://doi.org/10.1007/s11747-020-00749-9

Jarrahi, M. H. (2018). AI and work. *Business Horizons, 61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work. *Academy of Management Annals, 14*(1), 366–410. https://doi.org/10.5465/annals.2018.0174

Kizilcec, R. F. (2016). Transparency and trust. *CHI Proceedings*, 2390–2395. https://doi.org/10.1145/2858036.2858402

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). AI opacity. *Organization Science, 33*(1), 126–148. https://doi.org/10.1287/orsc.2021.1549

Lee, M. K. (2018). Algorithmic fairness. *Big Data & Society, 5*(1). https://doi.org/10.1177/2053951718756684

Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. (2015). Algorithmic management. *CHI Proceedings*, 1603–1612. https://doi.org/10.1145/2702123.2702548

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation. *OBHDP, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). AI resistance. *Journal of Consumer Research, 46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Murray, A., Rhymer, J., & Sirmon, D. G. (2021). Conjoined agency. *Academy of Management Review, 46*(3), 552–571. https://doi.org/10.5465/amr.2019.0186

Prahl, A., & Van Swol, L. (2017). Algorithm aversion. *Journal of Forecasting, 36*(6), 691–702. https://doi.org/10.1002/for.2464

Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and AI. *Journal of Marketing, 85*(1), 131–151. https://doi.org/10.1177/0022242920953847

Raisch, S., & Krakowski, S. (2021). Automation vs augmentation. *Academy of Management Review, 46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Seeber, I., et al. (2020). Machines as teammates. *Information & Management, 57*(2), 103174

Shrestha, Y. R., Ben-Menahem, S., & von Krogh, G. (2019). AI decision-making. *California Management Review, 61*(4), 66–83

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). AI in HR. *California Management Review, 61*(4), 15–42

von Krogh, G. (2018). AI in organizations. *Academy of Management Discoveries, 4*(4), 404–409

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Recommendations. *Journal of Behavioral Decision Making, 32*(4), 403–414