Enhancing Trust in Human-AI Collaboration: A Conceptual Review of Operator-AI Teamwork

Abduljaleel Hosawi[1], Richard Stone[2]
Dept. of HCI, Iowa State University Ames, USA[1]
Dept. of Industrial Engineering, Iowa State University Ames, USA[2]

Abstract—**Trust is vital to collaborative work between operators and AI. Yet, important elements of its nature remain to be investigated, including the dynamic process of trust formation, growth, decline, and even death between an operator and an AI. This review analyzes how the dynamic development of trust is determined by Team performance and its complex interaction with factors related to AI system characteristics, Operator competencies, and Contextual factors. This review summarizes current concepts, theories, and models to propose a genuine framework for enhancing trust. It analyzes the current understanding of trust in human-AI collaborations, highlighting key gaps and limitations, such as a lack of robustness, poor explainability, and effective collaboration design. The findings emphasize the importance of key components in this collaborative environment, including Operator capabilities and AI technology characteristics, underscoring their impact on trust. This study advances understanding of the nature of Operator-AI collaboration and the Dynamics of trust in trust calibration. Through a multidisciplinary approach, it also emphasizes the impact of Explainability, Transparency, and trust repair mechanisms. It highlights how Operator-AI systems can be improved through Design principles and developing Human competencies to enhance collaboration.**

Keywords—***Operator-AI collaboration; trust calibration; trust dynamics; explainability; transparency; trust repair mechanisms; cross-cultural trust; Clinical Decision Support Systems (CDSS); AI autonomy and influence; ethical considerations in AI; team performance; AI system characteristics; operator competencies; contextual factors; framework; limitations; robustness; Human-AI teaming; design principles; human competencies; predictability; reliability; understandability; over-reliance; automation bias; under-trust; trust measurement; trust erosion***

I. Introduction

The catastrophic failure of HAL 9000 in the movie “2001: A Space Odyssey” highlights the loss of trust between humans and artificial intelligence [1]. As humans cede more control to Al in vital areas, fostering and nurturing trust becomes critical. The line between beneficial cooperation and disastrous dependence is fine and requires careful consideration. The ongoing effort to achieve greater integration of Artificial Intelligence (AI) into complex operational environments is rapidly transforming industries and reshaping the nature of work. AI, can be defined as machines or systems capable of performing tasks that typically require human intelligence. For instance, the ability to learn, solve problems, perceive, and take decisions offers unprecedented potential for human, such as enhancing our efficiency, augmenting our capabilities, as well as tackling previously intractable problems [2]–[5]. However, to deploy and utilize these powerful technologies successfully, particularly in collaborative settings with human operators, must be on the establishment, maintenance, and calibration of trust [6]. As AI systems are increasingly guaranteeing autonomy and responsibility in high-stakes areas, such as healthcare, finance, transportation, manufacturing, and defense [4], [7], [8]. Gaining a deep understanding the dynamics of trust within operator-AI teams today is paramount not only for ensuring safety, but also effectiveness, and the realization of AI’s collaborative potential [9]–[11].

The concept of trust is multifaceted and dynamic in the context of human-AI interaction and has its own characteristics. Even though It’s drawing parallels with interpersonal trust, it still presenting unique challenges stemming from the AI’s non-human nature [6]. Generally, it is understood as an attitude or psychological state that involve a willingness to accept vulnerability (under conditions of uncertainty and risk) based on positive expectations regarding the AI’s future behavior and capabilities. That includes its predictability, competence, reliability, as well as the alignment with the operator’s goals [12]. There are three main aspects that are likely to affect this evolving attitude including the AI’s perceived performance. For example, observed reliability, accuracy, and consistency. The second aspect is process in particular understandability, transparency, predictability of its model. Third is the purpose and how far it aligns with operato’s goals, perceived benevolence, and integrity of its design [13], [14]. These three facets dynamically interact; for instance, if an AI’s reasoning is explainable (process transparency), may lead to better assessment of its competence by the operator and may also be easier in predictability (performance). That potentially fosters faster trust emergence or aids calibration after errors [15].

How the level of operator’s trust is calibrated in the capabilities of AI in a given context is critical, as it significantly impacts adoption behavior and overall team performance [16]. Distrust or (insufficient trust) could lead to the disuse (underutilization) or even abandonment of valuable AI assistance. Likewise, if an operator over-reliance or his/ her automation bias (excessive trust) results in issues such as complacency, reduced monitoring, and override AI errors or failure to detect them. This misuse, potentially, may have significant negative consequences [17].

According to Fig. 1, operators-AI collaboration could be best when the operator’s level of trust is appropriately aligned with the AI’ true capabilities. Failure to do so could lead to misuse by overtrusting the AI or under trust. For this reason, understanding trust dynamics (the entire lifecycle) is essential. In particular, how and when trust emerges, strengthened, and maintained before it decays or erodes. Trust can grow as a result of good behavior, and it can also be weakened as a

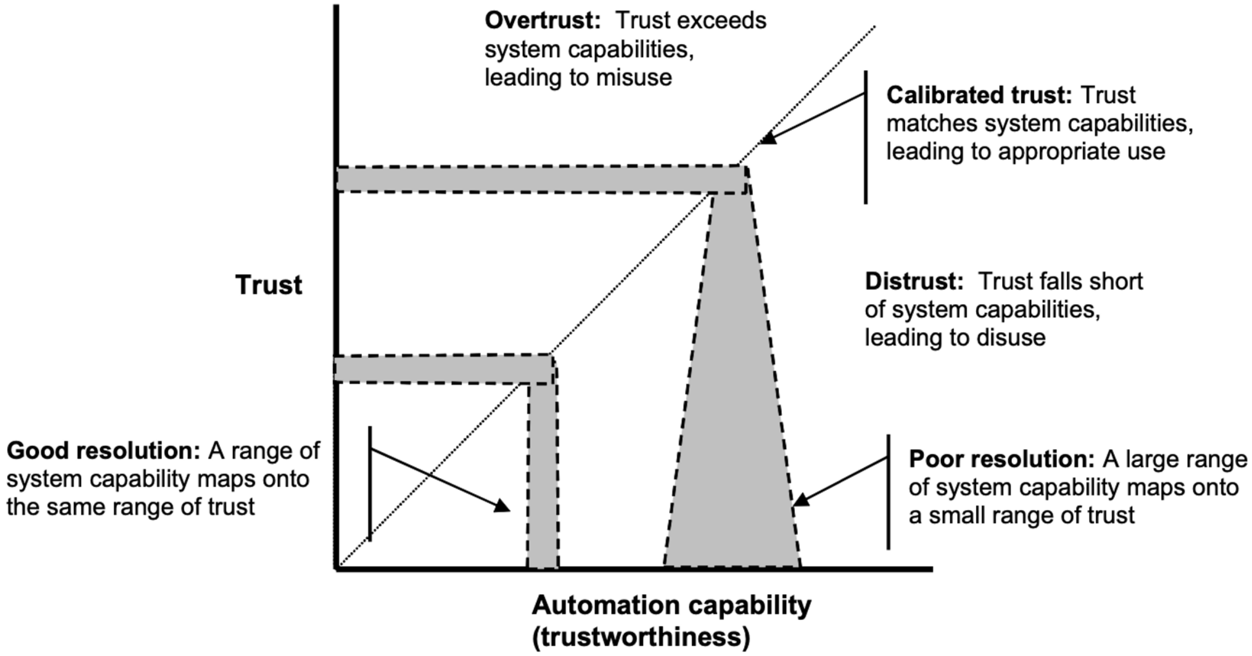
Overtrust: Trust exceeds system capabilities, leading to misuse
Calibrated trust: Trust matches system capabilities, leading to appropriate use
Trust
Distrust: Trust falls short of system capabilities, leading to disuse
Good resolution: A range of system capability maps onto the same range of trust
Poor resolution: A large range of system capability maps onto a small range of trust
Automation capability (trustworthiness)

Fig. 1. Calibrating operator trust according to the true AI capabilities. Note: This figure is taken from Lee and See [12].

result of errors and biases. In addition, knowing trust repair strategies when it is damaged is just as important as knowing its life cycle, and both are essential for designing systems that promote safe, effective, and appropriately calibrated human-AI collaboration [18].

The dynamic process of trust formation between the operator and AI throughout team performance is influenced by a complex interplay of factors related not only to the AI system, but also to operator, interaction and contextual factors. Factors related to the AI system include objective performance metrics, communication strategies employed by the AI, how the design of transparency and explainability are featured , and even AI's perceived cooperativity or embodiment [19]. Human operator factors encompass many significant elements such as individual differences including propensity to trust, domain expertise. In addition to cognitive biases, required competencies for interacting with AI, age, cultural background, and emotional responses [20], [21]. For the factors related to interaction and contextuality , literature suggests involvement in the nature of the task, the development of shared mental models, and team composition. Finally, the overall organizational environment suporting human-AI teaming [22].

Despite rapid growth and advancements in research interest to explore the trust dynamics in the Human-AI settings, significant gaps remain. There is still a persistent need for more longitudinal studies tracking trust dynamics over time [23], [24] , deeper exploration of cross-cultural variations in trust also still demanded [25], [26], validated approaches not only to measure trust but also to capture its dynamics and reveal its calibrated nature [27], [28], better understanding of trust erosion and repair mechanisms following AI errors, failures or bias [29], [30],continued examination of ethical considerations, (for example, regarding autonomy and influence) [21], [31] , and developing new generation of AI systems with true teaming competencies [8], [9], [32].

The present literature review aims to synthesize current knowledge on the dynamics of trus, its emergence, growth, and decay within operator-AI teams. Specifically, this review focuses primarily on new literature that reflects the latest advancements in the field. It examines the various types of AI systems relevant to this context of collaboration with human operators, explores the theoretical underpinnings of trust, analyzes the key factors influencing trust dynamics in addition to discussing strategies for building and maintaining trust. The review also investigates mechanisms of trust erosion and repair, and considers the implications for designing effective and trustworthy human-AI partnerships. Through consolidating research from diverse fields and highlighting persistent gaps, this review seeks to provide a comprehensive overview of the current state of understanding and inform future research directions. The aim is to contribute in enhancing trust and collaboration in the increasingly prevalent operator-AI teams.

A. Structure of the Review

The review is organized into several sections, as shown in Fig. 2: Section I: structured exploration of trust dynamics in operator-AI teams. Section II: classifying AI systems pertinent to operator-AI collaboration, and defining the scope of AI. Section III: primary theoretical frameworks informing trust, from different fields. Section IV: analysis of factors influencing trust dynamics. Factors categorized by AI, human operator, and context. Section V: temporal processes of trust (formation, calibration, decay, repair). Section VI: case study on Clinical Decision Support Systems (CDSS). Section VII: critical discussion and synthesis, challenges, limitations, unresolved questions. Section VIII: conclusion, core arguments, gaps, methodological approach, and future directions.

Thus, having explored the foundational dynamics of trust in operator-AI teams, it is now crucial to understand the diverse types of AI systems involved. Therefore, Section II will discuss classifying AI systems to identify the types relevant to operator-AI collaboration and define the AI scope in our review.

II. Defining the Scope: AI Systems in Operator-AI Teams

The term "Artificial Intelligence" encompasses a wide spectrum of technologies with different capabilities and functionalities [2]. In order to effectively discuss trust dynamics in operator-AI teams, it is crucial to specify the types of AI systems most relevant to collaborative contexts. Researchers need to ensure that interconnectedness and common goals exist or are possible [6], [9]. AI systems can be classified along several dimensions, including their capabilities relative to human intelligence, their core functionality based on how they process information, and finally, based on their intended role and scope within human interaction [3], [33]. Table I synthesizes common classifications relevant to operator-AI teaming. It's drawn from the original draft and sources including Jutel [2], Makarius [33], Bansal [34]–[36](Refer to Table I).

Table I [2], [33] illustrates AI systems classification in the context of operator-AI teaming. The table classifies the AI systems into three general groups: capability, functionality, and scope or role. Each group is divided further into categories, including Narrow AI (ANI), General AI (AGI), and Super AI (ASI), according to their capabilities. The second group contains AI systems classified based on functionality and encompasses Reactive Machines, Limited Memory (Alteration/Symbiotic under the Scope/Role classification), Theory of Mind, and Self-Aware. The last group is based on the scope or role of the AI model and contains six different types of AI systems. The review will focus on the Limited Memory AI systems as this category represents AI systems currently interacting with operators and possess sophisticated capabilities within specific domains. This category also can collaborate with humans to adapt and share goals with medium novelty ,content changing scope. Furthermore, these systems also can learn from experience, enabling advanced collaborations from the perspective of organizational integration and human roles [33]. Operator-AI teaming often involves tasks such as Augmentation, Alteration, and potentially Autonomous systems, where the relationship is complementary, symbiotic. The tasks also could involve careful monitoring and intervention by the human operator. The human role shifts from simply controlling a tool, coaching, or overseeing to collaborating with a true and more capable AI partner. In the future, Artificial General Intelligence (AGI) or Superintelligence (ASI) will need more examinations as they remain largely theoretical. Based on the above, this review focuses primarily on interactions with Narrow AI systems exhibiting Limited Memory functionality. These systems are more suitable for our study because they operate in roles that involve complementarity, symbiosis, or monitored independence (Augmentation, Alteration, Autonomous). In such systems, dynamic trust, reliance calibration, and effective teaming processes are most critical and currently researched. While future AGI might introduce radically different trust dynamics, investigating trust with current and near-future interactive Narrow AI could provide the necessary foundation. The principles discussed, particularly regarding transparency, reliability, communication, and human factors, are likely to remain relevant, even if the are modified as AI capabilities advance.

Having established the scope and types of AI systems in this context, it is time to understand the theoretical foundations for studying trust. Therefore, Section III delves into the primary theoretical frameworks drawn from various fields that inform our understanding of trust in the context of human-AI collaboration.

III. Theoretical Foundations of Trust in Human-AI Collaboration

Several theoretical frameworks represent a starting point to understand trust dynamics in operator-AI teams.

1) First, Social Exchange Theory (SET): Originating in sociology [22], according to SET, social relationships are built on the exchange of resources. Individuals seeking to maximize benefits and minimize costs. Human operators are more likely to cooperate if the perceived benefits of teaming with AI (e.g., improved decision making, reduced time, increased efficiency) outweigh the costs (e.g., risk of errors, loss of control, reduced cognitive effort). Trust develops when the perceived benefits consistently outweigh the costs, and enhances willingness to continue cooperation (reliance) with the AI [37], [38]. Several factors such as perceived fairness, reliability, and competence of the AI can influence this cost-benefit analysis and subsequent trust [39]. However, applying SET directly faces various challenges due to the AI's lack of social factors and the opacity of its "intentions" [40].

2) Cognitive Load Theory (CLT): Focuses on the limitations of human working memory, and distinguishes between three types of memories: intrinsic load (task complexity), extraneous load (interface/information presentation), and germane load (schema construction/learning). CLT developed in educational psychology Sweller. In the context of human-AI interaction, CLT states that a poorly designed AI systems can impose high extraneous cognitive load (e.g., confusing interfaces, overly complex explanations). This can hinder understanding and trust formation [41], [42]. Conversely, well-designed AI can reduce cognitive load by automating routine tasks or presenting information clearly and concisely. A well-designed AI model more likely help in freeing up cognitive resources for operators to better evaluate the AI's output and build trust [43]. Managing cognitive load is thus crucial for effective collaboration and trust [44].

3) Human-Computer Interaction (HCI) and Technology Acceptance Models: HCI research provides numerous models for understanding user interactions with various technologies.The Technology Acceptance Model (TAM) and its more recent extensions such as "AI Adoption TAM", "User Acceptance of AI Systems", or "Trust and Technology Acceptance Models for AI" are some of those basic models. These models emphasize perceived usefulness and perceived ease of use as key elements that determinants of technology acceptance. Both elements

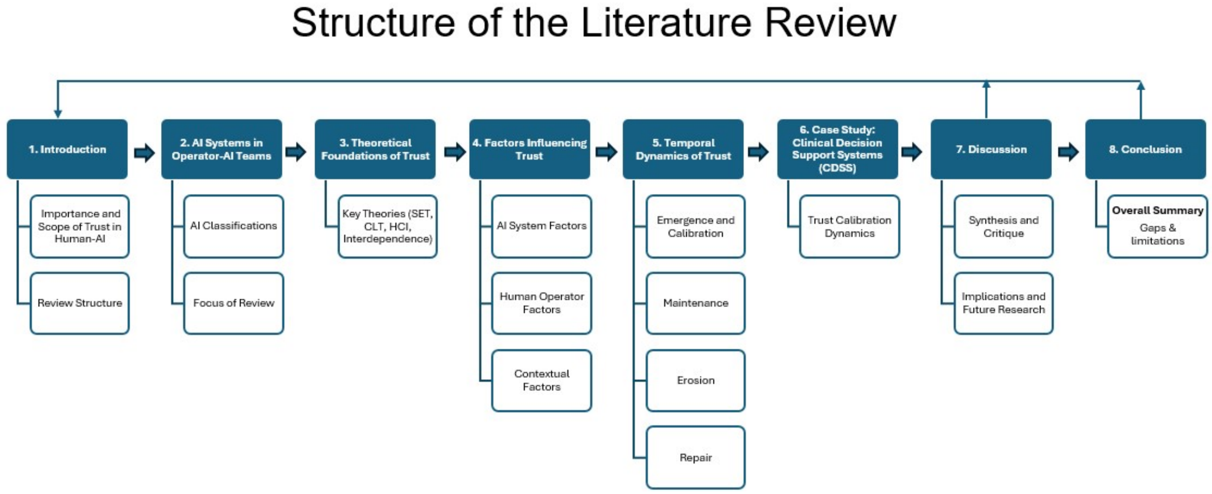
Structure of the Literature Review
1. Introduction
Importance and Scope of Trust in Human-AI
Review Structure
2. AI Systems in Operator-AI Teams
AI Classifications
Focus of Review
3. Theoretical Foundations of Trust
Key Theories (SET, CLT, HCI, Interdependence)
4. Factors Influencing Trust
AI System Factors
Human Operator Factors
Contextual Factors
5. Temporal Dynamics of Trust
Emergence and Calibration
Maintenance
Erosion
Repair
6. Case Study: Clinical Decision Support Systems (CDSS)
Trust Calibration Dynamics
7. Discussion
Synthesis and Critique
Implications and Future Research
8. Conclusion
Overall Summary
Gaps & limitations

Fig. 2. Organization of the literature review.

are essential and may be considered related to trust more recent models, including the Unified Theory of Acceptance and Use of Technology (UTAUT), incorporate additional factors. Some important factors are social influence and facilitating conditions [45], [46]. In the field of HCI, multi-dimensional models of trust in automation and AI have emerged (e.g., competence, benevolence, encompassing, integrity, predictability, and dynamic). These models highlight the importance of system characteristics such as performance, and design, user characteristics such as disposition to trust, and finally contextual factors in shaping trust over time [42], [47], [48].

4) Theory of Mutual Dependence or Interdependence Theory: Based on these theories, interdependence among team members is crucial for achieving common goals. This applies whether this interdependence is within traditional teamwork between humans or between humans and artificial intelligence [49]. In the team, trust is cornerstone and can act as an auxiliary worker in interdependent relationships, facilitating coordination, communication, and willingness to rely on others [50]. In human-AI teams, this interdependence should go both ways, requiring operators to trust the AI's capabilities within its designated role, while the AI (through its design) must reliably fulfill that role [41]. Correctly calibrating trust dynamics in light of AI capabilities is crucial and a significant challenge, especially given the impact of influential factors. Also, it's important to establish shared awareness and mental models between human and AI agents [51]–[53]. These frameworks collectively highlight that trust in operator-AI teams is a complex phenomenon influenced by several factors such as perceived utility, cognitive feasibility, system design, user psychology, and the dynamics of interdependent collaboration.

These diverse theoretical perspectives provide a robust foundation for understanding trust in human-AI contexts. We can now turn to a detailed analysis with these theoretical frameworks established. Section IV systematically analyzes the multifaceted factors identified in the literature that influence trust dynamics.

IV. Factors Influencing Trust in Operator-AI Teams

Trust is not static but emerges and evolves based on interactions between the operator, the AI system, and the context. Key influencing factors can be categorized as follows:

A. AI System Factors

1) Performance and Reliability: Performance is arguably the most fundamental factor in building trust. It must be consistent, accurate, and reliable [12], [39], [42]. Predictability, or the extent to which the AI behaves as expected, is also crucial [40], [54]. In addition, errors or inconsistent performance significantly erode trust between operators and AI systems [24], [29].

2) Automation and Trust: Fig. 3 illustrates a model of trust in automation [12], [55], indicating the factors that directly impact how a human operator builds and maintains trust in automated systems and how that trust influences the operator's reliance on automation. At the top of the figure, the large box indicates the general context associated with the formation

TABLE I. Classifications of AI Systems Relevant to Operator-AI Teaming

Classification Axis	Category	Description	Examples Relevant to Teaming	Typical Human Role
Based on Capability	**Narrow AI (ANI)**	Designed for specific tasks; cannot perform outside training domain.	Facial recognition, Driving assistance, Diagnostic aids, NLP chatbots, Go programs	Controller, Collaborator
	General AI (AGI)	Possesses human-like cognitive abilities; can learn and perform various intellectual tasks (mostly hypothetical).	Hypothetical versatile assistants or partners.	Co-creator, Comprehend
	Super AI (ASI)	Surpasses human intelligence across all domains (purely hypothetical).	Speculative superintelligences.	Comprehend
Based on Functionality	**Reactive Machines**	Reacts to stimuli; no memory of past experiences.	Early game AI (e.g., Deep Blue).	Supervisor (limited teaming)
	Limited Memory	Learns from past data/experience to inform future decisions; most current AI applications.	Chatbots, CDSS, Recommendation systems, Autonomous vehicles.	Controller, Collaborator, Coach
	Theory of Mind	Understands emotions, beliefs, intentions (conceptual/developing).	Advanced virtual assistants, Social robots.	Collaborator, Co-creator
	Self-Aware	Possesses consciousness, sentience (theoretical).	Hypothetical conscious AI.	not specified
Based on Scope/Role	**Automation / Substitution**	AI replaces human tasks; low novelty, content-changing scope.	Assembly line robots, Basic DSS.	Controller
	Amplification / Supplementary	AI enhances human analysis; low novelty, context-changing scope.	Predictive analytics tools.	Conductor
	Augmentation / Complementary	AI assists humans in complex tasks; medium novelty, content-changing scope.	Surgical robots, Advanced DSS, Go programs, Chip design AI.	Collaborator
	Alteration / Symbiotic	AI and humans co-create/adapt; medium novelty, context-changing scope.	Deep learning systems needing expert interaction.	Co-creator
	Autonomous / Independence	AI operates largely independently; high novelty, content-changing scope.	Self-driving vehicles (higher levels).	Keep-in-Check (Monitor)
	Authentic / Singularity	AI potentially surpasses human roles; high novelty, content-changing scope (hypothetical).	Superintelligence (ASI).	Comprehend

Note: Capability and Functionality classifications adapted from Jutel [2], common AI taxonomies, scope and role classification adapted from Makarius [33].

and reliance of trust, including the individual's characteristics (such as experience and personality), the organization's organizational structure (such as policies and procedures), culture (differences in shared norms and values regarding trust in technology), and the work environment context (the circumstances and situations in which automation is used).

In the same box, the figure shows the factors that influence the level of automation implementation, which in turn contributes to the operator's perception and level of trust. For example, these factors include what others say, rumors, and interface design quality. Other factors include workload, effort required to participate, self-confidence and personal capabilities, perceived risks, and system setup errors related to automation. In the middle of the model, four boxes appear from the left to the right, representing the mechanism through which the trust audit process progresses, from its creation to its adoption. This process begins with gathering information, perceptions, and beliefs about automation, followed by an evaluation by the operator in preparation for transforming them into practical actions by determining how to deal with the automation and, finally, whether to adopt it. Feedback between some stages may alter perception, evaluation, and the operator's readiness to embrace or avoid the automation. This development occurs through dynamic interaction with the application context and influencing factors.

In the lower right of the model, we find the automated system and its related capabilities and usage procedures. On the opposite side (bottom right), the model depicts the user interface and its impact, including ease of use, information reception and clarity, and ease of feedback. In the middle part, the model demonstrates that the operator receives information about the automation that may contribute to their understanding and confidence, such as the purpose of the automation, how it is implemented, and the quality of performance (refer to Fig. 2).

3) Transparency and Explainability: Explainability, the ability of human operators to understand how an AI arrives at a decision or recommendation, and having insight into its processes, true capabilities, and limitations (transparency) are critical for building calibrated trust. Transparency is vital, especially for complex or "black box" models [56], [57], [56] ExplainableAI. Explainable AI (XAI) is a technique that aims to provide transparency insight, which is particularly important in high-stakes domains [58], [23], [30]. Even the availability of explanations, whether accessed or not, can increase perceived trustworthiness [24], [59].

4) Interface and Interaction Design: Several essential elements, including User-friendly interfaces, precise feedback mechanisms, and intuitive controls, contribute significantly to perceived ease of use and reduce extraneous cognitive load, fostering positive experiences and trust [6], [40], [47]. Human-centered design (HCD) principles are critical and emphasize

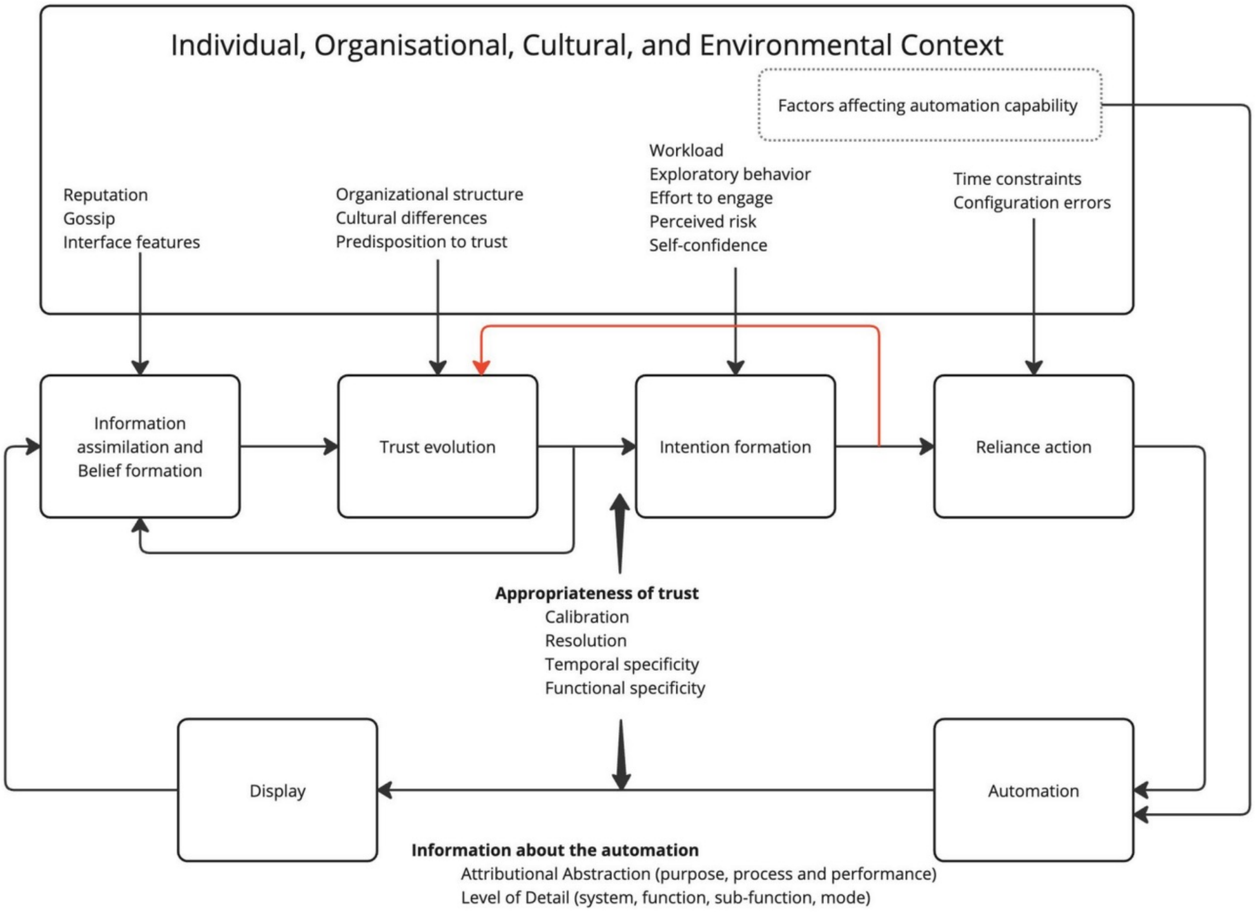

Fig. 3. The model of trust in automation, as originally proposed by [12] and modified by Sheridan [55]

tailoring the AI system to operator needs, capabilities, and context [6], [40]. Anthropomorphism refers to the design of AI with human-like characteristics (e.g., appearance, communication style) can sometimes foster emotional trust and rapport. However, it can also lead to unrealistic expectations or uncanny feelings if not implemented carefully [37], [40], [60].

B. Human Operator Factors

1) Disposition to Trust: Individuals vary in their inherent propensity to trust others or even to reassure to use technology, influenced by personality traits and past experiences [37], [42]. Experience and Expertise: Prior experience with AI or similar technologies, as well as domain expertise, are significant factors and can shape mental models and expectations of AI systems, influencing trust levels [61], [62]. Training can help develop accurate mental models but must be carefully designed [43], [63].

2) Cognitive Biases: Operators are susceptible to biases like automation bias (over-reliance on automation) or confirmation bias, leading to miscalibrated trust [6], [43]. Cultural Background: Cultural norms and values can significantly influence expectations, communication styles, and perceptions of technology, leading to cross-cultural differences in trust towards AI (GSD Venture Studios, n.d.), [26], [40]. AI design and interaction must consider these cultural nuances [26].

C. Contextual and Task Factors

1) Task Complexity and Criticality: Trust requirements in the operator-AI settings are high and often higher for complex tasks or in high-stakes environments, especially where the consequences of failure are severe (e.g., healthcare diagnostics, autonomous flying, military operations) [42], [64], [65].

2) Risk and Uncertainty: The capabilities of AI systems vary depending on the model and its characteristics. Higher perceived risk or uncertainty in the operational environment often necessitates greater trust in the AI's ability to perform reliably [62], [66].

3) Organizational Culture: The organizational context, including policies regarding AI use, transparency norms, and support structures, can influence operator trust [67].

Understanding these categories of influencing factors is essential, but trust is also a dynamic process that changes over time. Therefore, Section V focuses on the temporal processes of trust, examining how it initially forms and is calibrated and maintained over time.

V. Temporal Dynamics of Trust: Emergence, Calibration, Maintenance, Erosion, and Repair

Trust in an operator-AI team is dynamic rather than static. It is created and evolves over time, up and down, due to ongoing interactions with surrounding factors [68]. Operators assess trust and interpret AI behavior in specific contexts. Understanding the lifecycle of trust is crucial: how it is created, grows, or erodes due to adverse events, the strategies for repairing it when it is damaged, and how it is calibrated and maintained. The above elements are essential for developing strong and resilient partnerships in human-AI teams.

A. Trust Emergence and Initial Calibration

Initial trust between the operator and the AI is formed before or during early interactions [12]. Usually, it is shaped by factors such as relying on heuristics, pre-existing attitudes, and surface-level characteristics rather than deep experience. Other factors that may influence initial trust include the operator's disposition to trust technology in general [69], preconceptions based on reputation or prior experiences with AI or similar systems, the perceived purpose and design pedigree of the AI (e.g., developed by a reputable source), in addition to interface characteristics that signal professionalism or usability [12]. Initial interactions are critical; early positive experiences demonstrating reliability and competence can establish a positive trajectory. Similarly, early failures can disproportionately damage nascent trust [12], [69].

Operators begin trust calibration as the interactions with the AI progresses. They adjust their level of trust to align with the AI's perceived capabilities and limitations in the context [12]. AI processes and boundaries require transparent and accurate feedback on their performance to succeed in the calibration effectively. Providing explanations, particularly those relevant to the task and understandable to the operator (e.g., example-based or counterfactual explanations according to the contexts), can aid this process by supporting operators in building more accurate mental models of the AI's reasoning [6], [16]. However, ineffective or overly complex explanations can do the opposite. Poor explanations can hinder calibration. Explicitly revealing AI uncertainty or confidence levels can also be profitable in aiding calibration, even though it may sometimes decrease overall trust levels [6], [70]. Several models are being explored and incorporating Bayesian learning, which refers to applying probability theory to learning from data, especially regarding trust and interpretability [71]. The way operators update their appraisal of AI capabilities could be captured through the exploration based on observed outcomes [72] . To avoid the pitfalls of over-trust or under-trust, successful calibration is essential.

B. Trust Maintenance

The process of maintaining trust begins once it is established and calibrated between the operator and the AI [42]. Sustained reliability and predictable performance are crucial [12], [42]. Operators can anticipate AI actions and integrate them smoothly into their workflow through predictability. [54]. The importance of transparency goes beyond just the initial calibration but also effective communication, and ongoing verification, understanding. This importance is vital in a changeable environment or when AI's behavior is modified. Helpful strategies, such as proactive updates or confirmations, can reinforce perceptions of reliability and teamwork, contributing to trust maintenance [70]. Furthermore, AI systems designed with intuitive human-centered principles that align with operator workflows and cognitive limits are crucial and more likely to sustain trust over time by ensuring good user experiences [6], [11]. Sharing learning, an environment where both humans and AI can exchange knowledge and experience through collaboration, strengthen relationships, and preserve trust [11].

C. Trust Erosion

Trust can be described as "hard-won but easily lost". It has several actors that lead to trust erosion: First, AI errors and failures are the most direct reasons trust decreases. Although trust can be easily affected be errors, the type of error, its frequency, and its severity may have the greatest impact [12], [42], [54]). Sometimes, occasional errors, especially critical or unexpected ones, can significantly affect or damage trust.

The next cause can be the lack of transparency and explainability [57], as operators need to know everything while collaborating with their intelligent partners. Issues can arise when AI fails or behaves unexpectedly. Black-Box issues can be a real problem by hindering users' understanding of the cause of system errors. As a result, users may feel frustrated, uncertain, confused, and lack control [12], [16], [73]. Also, poor interpretations can lead to similar harms and results, leading to dissatisfaction or overtrust of the system.

Unreliable performance can make predictability difficult [42], negatively impacting the operator's trust in the system. For instance, surprises in AI system behavior can cause system performance to fluctuate, leading to a gradual erosion of trust over time [12], [63]. Other harmful reasons are perceived bias and unfairness. AI biased or unfair actions would cause damage to trust, particularly if they negatively impact the operator or specific groups. Trust can be severely damaged [33], [63]. Similarly, a discrepancy between operator expectations and the actions of an AI system can negatively impact trust between the two. Sometimes, AI capabilities are exaggerated, leading to feelings of disappointment or distrust, especially in early interactions with the system [12]. Trust can also be decreased as a result of poor usability or communication and can lead to trust decreased [12]. Communication is vital for operators while teaming with an AI partner. Awkward interfaces, ambiguous feedback, or ineffective communication protocols may create friction, confusion, and frustration. Leading to negative impacts on the user experience and indirectly eroding trust [12], [16], [70]. Finally, the context also matters because errors in high-stakes situations typically cause more significant trust erosion than errors in low-stakes contexts [42], [73].

D. Trust Repair

Trust may be easily damaged and must quickly be repaired [74]. Rebuilding trust after damage is not an easy challenge in human-AI interaction. Arguably, it can be even more so than in human relationships due to the AI's lack of genuine social recourse (e.g., repentance, sincere apology, remorse). Research indicates that simple apologies or explanations for

AI failure have shown only marginal effectiveness in restoring trust [29]. The study found apologies and explanations to be the most effective relative strategies compared to not taking action or no repair; however, their overall impact was not very influential.

First, acknowledging and explaining failures is a must action. Transparency about the source of an error is crucial, moving beyond simple apologies [6], [16]. Tailored explanations that address the specific failure context may be needed. Also, demonstrating corrective action could be a good strategy [74]. We need to confirm to the operators that the underlying cause of the failure has been handled (e.g., communication, policy, algorithm update, retraining) and that affected trust can be rebuilt faster. Nevertheless, It is essential to work towards consistent future performance. Reestablishing a track history of reliable and predictable behavior after a failure and overwriting the negative experience can be a strategy for repairing trust [6], [12]. Additionally, an innovative solution could be providing a means of control to the operator or the appropriate agency as support; this may help restore damaged trust [75]. This support could provide transparency and regain trust. Allowing operators to understand, verify, and potentially override AI actions, such transparent procedure can aid in restoring a sense of control and mitigate the vulnerability associated with trust decrease [31]. Longitudinal studies demonstrate that explanations accelerate trust restoration after a malfunction. This acceleration can be beneficial even if not always accessed [24]. However, We must acknowledge a need to develop accurate and effective solutions that guarantee systematic trust repair strategies for human-AI teams. This demand remains an open research area.

E. Trust Measurements

Trust is multifaceted and changes over time, so no single metric may be sufficient to assess it or measure the performance of the operator and AI team. Table II and Table III summarize the most important metrics used to assess trust and evaluate the effectiveness of the operator and AI team.

Table II shows several metrics for measuring trust drawn from various fields (human factors, HCI, and psychology research). Questionnaires and surveys are the most commonly used metrics due to their ease of use and are a branch of self-report metrics. Standardized trust scales are also used to assess predictive ability, perceived reliability, and trustworthiness, particularly in automation and artificial intelligence. There are two other questionnaire and survey scales: Likert-type and semantic differential scales, which ask the user to rate their level of agreement with a specific statement and use bipolar adjectives such as trustworthy-untrustworthy, respectively. The final scale used in questionnaires and surveys is Interview/Focus Groups, which focus on identifying causes and explaining phenomena, such as which factors specifically led to a high or low level of trust. The second type of scale that can be used to measure trust is Behavioral Measures (objective), which focus on observing how the operator interacts with the artificial intelligence system. Three measures under this category are Reliance/Adherence, Task Performance, and Interaction Patterns. Each has its purpose and nature; some are used for measuring the frequency of a particular phenomenon, others for error rate or the extent of the operator's need to search for information such as clarification or assistance. The final category of trust metrics that can be employed in an operator-AI team are Physiological Measures (Objective but more complex). These metrics involve observing a human operator's vital signs, such as Heart Rate Variability (HRV), Galvanic Skin Response (GSR)/ Electrodermal Activity (EDA), and Eye Tracking.

Table III shows which metrics are used for measuring operator-AI teamwork. Operator-AI Collaboration is a crucial measure, and its metrics overlap with trust metrics [76], but they focus on the effectiveness and nature of joint work. Operator-AI Collaboration is measured using the Team Performance Metrics, which measure efficiency, accuracy, and workload—interaction Quality Metrics for evaluating communication effectiveness, mutual understanding, role clarity, allocation, and adaptability. For Subjective Collaboration Measures, usually asks perators to reate some aspects. For instance, "I felt the AI was a bad manager" or "I comprehended how the AI was finishing the task". In addition, Subjective Collaboration Measures could also be used to find out how well the operator feels integrated with the AI as a team (refer to Table III).

The important question is what is the appropriate measure for measuring trust between operators and artificial intelligence (AI)? In addition, what is the best way to assess their collaborative relationship? Given the fluid and multifaceted nature of trust, more than one type of measure may be needed depending on the scope of the application and its extent, which underscores the importance of context. Trust must be measured at various levels for the trust enhancement model proposed following this review, which will primarily focus on enhancing trust in the operator-AI environment. There will likely be a need for subjective (e.g. questionnaires) measures with objectives (e.g., behavioral) related to the human element, the AI system, and context. These measures are also essential in most fields, such as ethical principles, measuring AI's true capabilities, and the human operator's readiness. By integrating these measures, weaknesses that have arisen in previous studies due to the absence of necessary measures can be overcome. Trust formation, decay, and repair mechanisms are fundamental, especially in high-stakes environments. To illustrate these concepts in practice, Section VI presents a case study focused on Clinical Decision Support Systems (CDSS). The CDSS provides a window to examine the evolution of trust through its various stages.

VI. Case Study: Trust Calibration Dynamics in Clinical Decision Support Systems

Many challenges of trust dynamics are uncovered in the CDSS case study [77], particularly in the calibration. These challenges and the impact of system design are vividly described in the context of Clinical Decision Support Systems (CDSS). In a high-stakes environment where decisions require high accuracy and reliability (examination and prescribing chemotherapy to patients), the study examined the impact of XAI's explanations on trust levels among 41 medical practitioners. Four categories of explanations were compared (example-based, global information-based, local context-based, and unrealistic explanations). Participants interacted with scenarios featuring both correct and incorrect AI recommenda-

TABLE II. Metrics for trust in operator-AI collaboration

Metric Category	Measurement Methods	Focus of Measurement
Self-Report	Questionnaires/Surveys, Interviews/Focus Groups	Subjective perceptions and beliefs about AI trustworthiness.
Behavioral	Reliance/Adherence, Task Performance, Interaction	Observable actions and patterns of operator behavior when interacting with AI.
Physiological	HRV, GSR/EDA, Eye Tracking	Physiological responses that may correlate with trust, stress, or cognitive load related to the AI.

TABLE III. Metrics for measuring operator-AI collaboration

Metric Category	Measurement Methods	Focus of Measurement
Team Performance	Efficiency, Accuracy, Workload (Subjective/Objective), Situation Awareness	Effectiveness of the combined human-AI system in achieving task goals.
Interaction Quality	Communication Effectiveness, Mutual Understanding, Role Clarity/Allocation, Adaptability	Nature and quality of the interaction between the operator and the AI system.
Subjective Collaboration	Collaboration Experience Questionnaires, Perceived Team Cohesion	Operator's perceptions and feelings about working collaboratively with the AI.

Note: Tables II and III are adapted from [76]. Table II focuses on trust measures, and Table III shows the measures used to evaluate team performance.

tions, then trust was assessed using self-reported cognitive dimensions (understandability, reliability, competence) and behavioral indicators (agreement with AI, switching decisions, overall human-AI performance).

1) Trust Emergence and Understandability: The case study results indicated the importance of the type of explanation. The study found that the initial impressions were significantly influenced by the type of explanation provided. Example-based and unrealistic explanations were easier to understand and distinguish than other explanations (public and local). According to the results, explanations that included concrete examples or "what if" scenarios were more straightforward to comprehend than abstract scores. This finding aligns with psychological principles suggesting humans favor familiar, causal explanations. In addition, some participants have to separate local and global interpretations or misinterpret them due to difficulty in understanding them, ultimately hindering the possibility of these interpretations strengthening initial trust. This finding is consistent with psychological principles that humans favor familiar, known causal explanations. Furthermore, some practitioners have been forced to misuse local and global explanations because they are difficult to understand, which ultimately leads to these explanations not strengthening initial trust. This finding suggests that humans prefer only well-known logical explanations.

2) Trust Calibration Challenges: While providing explanations (regardless of class) improved overall human-AI task performance compared to no explanation, a critical finding emerged regarding calibration: explanations did not significantly help practitioners recognize incorrect AI recommendations. Explanations did not help practitioners substantially recognize incorrect AI recommendations. Participants found defective AI suggestions, including the no-explanation baseline, not easy to disagree with across all explanation conditions. Furthermore, providing explanations significantly increased participants' agreement with AI recommendations compared to not providing any explanations. Therefore, suggesting explanations may inadvertently lead to overreliance or confirmation bias. Participants may have used explanations heuristically (System 1 thinking) as a signal of AI competence rather than engaging in deeper critical analysis (System 2 thinking), especially given workflow pressures.

3) Trust Decay and Contextual Needs: Qualitative feedback revealed critical factors that could lead to trust erosion or prevent trust from developing appropriately, even with explanations. Participants provided feedback expressing their concern that in their feedback, participants voiced concern when explanations seemed disconnected from clinical context or task constraints (For example, explanations that are unrelated to the patient's condition or seem unrealistic). In addition, participants highlighted the need for assurances regarding the validity and capability of the explanation method (e.g., data integrity used, consistency) and allowing for desired tailoring and customization to fit professionals' specific workflow and information needs. The lack of multi-step explainability or the ability to ask follow-up questions about an explanation was also a barrier to deeper understanding and trust validation.

This critical case study powerfully demonstrates that simply providing explanations is insufficient for ensuring appropriate trust calibration in operator-AI teams. The study shows that explanations' type, design, usability, and contextual relevance critically influence how operators perceive, understand, and ultimately rely on AI systems. It highlights the risk of explanations fostering over-reliance if not carefully designed and integrated. In addition, the findings underscore the need for interfaces that support critical thinking, address usability constraints, provide necessary assurances, and allow for user adaptation and deeper inquiry. This case study illustrates the complexities of confidence calibration in a credit decision support system. Based on these specific examples, we now turn to Section VII, which provides a critical discussion and summarizes the key findings from all the previous sections.

VII. Discussion

AI teams' evolving and expanding landscape requires operators to understand the dynamics of trust, a critical factor in successful collaboration. This literature review explores this vital topic, highlighting the shift toward viewing AI as a team partner rather than simply a tool, revealing the complex nature of trust in these contexts, synthesizing various theoretical perspectives and research findings, and revealing key themes, significant challenges, and critical tips for future work. Building upon this overview, the following section delves into synthesizing the key findings regarding the multi-determined nature of trust dynamics.

A. Synthesis: The Multi-Determined Nature of Trust Dynamics

Researchers have consistently demonstrated that trust has a lifecycle through which it emerges, grows, and begins to decay. Trust is not driven by any single factor. Instead, it results from a complex interplay between the AI system's characteristics, the human operator's attributes, and the interaction context [12], [42]. AI performance and reliability are the foundation of trust [12], [37]. But these important factors are just as important as other key factors, such as transparency and explainability. This importance is particularly true in high-risk environments or sensitive systems [6], [16], [73]. Clear explanations facilitate operator roles and help operators calibrate trust; in other words, while performance and reliability form the foundation of solid trust, it is important to remember that "reliability" is not a fixed measure. Its perception can be highly subjective. An AI system may be objectively reliable, but trust can quickly erode if the operator fails to make decisions or if occasional errors are not explained and addressed effectively. This finding suggests that subjective perceptions of reliability, influenced by transparency and interpretability, are just as important, if not more so, than objective technical performance. On the other hand, complex or poorly designed explanations should be avoided. They can hinder usability, cause serious problems, and increase cognitive load. Poor explanations can also lead to over-reliance on AI or avoiding it over time.

While AI capabilities form a foundation for trust, the role of the human operator is equally vital. The human element is equally critical. Individual differences, willingness to trust, expertise, cognitive styles, and competencies, all these factors affect how operators perceive AI [12], [20], [78]. Cultural background emerges as vital in trust dynamics, nevertheless often Ignored. An operator's artistic background is an essential factor that can significantly influence trust perceptions and expectations of the team in trust [26]. The frequent neglect of cultural and even artistic backgrounds in current research is significant. This gap may be due to the predominant focus on the technical aspects of AI. However, the fundamental truth is that this technology ultimately approaches humans from an inherently cultural and individual perspective. For example, an operator from a particular culture may prioritize AI's ability to integrate with group systems seamlessly. In contrast, an operator from an individualistic culture may emphasize security and privacy more. Similarly, an artist may be more attuned to AI's creative potential and fine-grained judgment, while an engineer may prioritize precision and logical reasoning. These differences can lead to significant variations in trust formation and team dynamics. Developing shared mental models and how cooperative they seem can greatly affect the teamwork and trust levels [27], [22], [70]. The Hanabi experiment highlights a crucial point. The game experiment highlights how subjective perceptions of teamwork and agent interpretability can diverge from objective performance. Sometimes, rule-based agents are preferred over objectively comparable, but less predictable, learning-based agents [27], [79]. In other words, people such as operators might choose to work with an AI that they "understand" and "feel comfortable" with, even if that AI isn't the one that is better or will not necessarily achieve the highest score. Trust and perceived teamwork matter. The way in which an AI agent is predictable and interpretable can influence them.

Although the above synthesis presents a clear picture of trust dynamics and the factors influencing them, it is important to recognize that there are still many ongoing challenges and limitations in the operator-AI environment that merit further discussion.

B. Critique: Persistent Challenges and Limitations

Despite the advances, several critical challenges continue in operator-AI settings. Firstly, the problem of trust calibration stays central as indicated by Naiseah and Vössing in [16] and [6]. According to Naiseah and Vössing, even well-intentioned transparency mechanisms, such as explanations, don't automatically guarantee appropriate reliance; explanations can sometimes lead to over-trust or fail to help users detect AI errors. A significant limitation is designing adequate interfaces and precise interactions that promote critical evaluation (System 2 thinking) rather than heuristic acceptance (System 1 thinking).

Another significant issue is that we are still unable to understand the deep problem of trust repair, especially after significant AI failures, and it's still challenging to achieve it successfully. Current strategies have not gone beyond Apology, and basic explanations are still insufficient [29]. One of the challenges that complicates repair efforts is the asymmetry of the operator-AI relationship. AI lacks genuine remorse and social accountability, which is taken from interpersonal trust literature. Additional research is needed on robust repair mechanisms. The new studies potentially involve verifiable demonstrations of learning and correction by AI.

Researchers conduct most studies in AI from a Western perspective, often ignoring other perspectives. Furthermore, much of the researches are conducted in laboratories and controlled environments. Often, non-experts participate, or participants perform simple roles. For instance, while useful, the Hanabi game [27] may not capture high-stakes complexities. Longitudinal studies can help us better understand trust dynamics, especially when done in real-world settings. These studies allow us to track changes and influencing factors over time. These studies are rare, but they are essential [24], [59]. Furthermore, a significant portion of research on AI technologies originates from Western cultural contexts. This central view can limit the generalizability of findings and potentially lead to culturally inappropriate AI designs [25], [26]. Beyond these practical challenges, broader cultural issues also complicate the field.

Fourth, the very definition and measurement of "teaming" and "trust" in the human-AI context are still evolving, which confirms the need for more extensive and accurate studies. Metrics that primarily focus on measuring task performance fail to measure the subjective quality of the interaction, potentially impacting long-term adoption and effectiveness [27], [79]. There is still a great need for vital multidimensional measures that can capture tangible cooperation, shared importance, and measured dependence. Given these challenges, the findings compiled in this review have direct implications for AI design, training, and policy.

C. Implications for Design, Training, and Policy

The synthesized results of our review confirm the direct implications of trust on AI design, particularly concerning human-centered approaches [6], [58]. The repercussions mean prioritizing not just AI System performance but also designing transparency and explainability features that are genuinely useful and understandable to the specific operator in their context [6], [16], [30]. To achieve this, it is required to involve end-users in the design process and considering their needs and preferences. As well as cognitive load, usability, and design for effective communication [70]. Incorporating mechanisms for behavioral synchrony [19] could also enhance teaming. Trust dynamics also significantly impact how tasks are allocated between humans and AI. When operators do not trust AI, they often rely on manual methods, decreasing efficiency. Conversely, over-reliance on AI can lead to unjustified trust and potentially fatal errors. Therefore, understanding and managing trust goes beyond just a psychological issue; it is a fundamental aspect of effective work design in operator-AI teams. In addition, considering the division of labor and interaction patterns from an organizational design perspective can provide structure [11], [41]. Crucially, designs must be sensitive to potential cultural differences in expectations and interaction styles.

For operator training and competency development, the focus should be on the key factors that enhance operator trust. In particular, there is a demand to focus on the key appropriate mental models of AI capabilities. Critical evaluation skills are essential for operators; it is important for training to go beyond technical skills. Operators need "trust literacy", meaning they can critically evaluate AI, calibrate their trust appropriately, and understand when and how to intervene. In addition, operators must develop the competencies necessary to interact with AI, including digital literacy, adaptability, collaboration, and problem-solving skills [78]. Finally, training should explicitly address foundational issues such as automation bias and calibration strategies. Such capabilities have implications for curriculum development and assessment across multiple disciplines.

Form organizational policy, promoting and fostering a new culture that supports transparency, ethical oversight, and learning from successes and failures in human-AI teaming is vital [6], [33]. As a necessary direction for policies for the manufacture and employment of human-based artificial intelligence, clear guidelines on roles, responsibilities, and accountability within human-AI teams are essential, especially in high-stakes domains [6], [73]. Who is responsible when something goes wrong in a human-AI team? The dynamics of trust can blur the lines of accountability. Who bears responsibility if an operator gives AI too much trust and follows incorrect advice? The operator or the AI designer? These legal and ethical considerations need to be studied and regulated. Policies should also consider the potential for differential impacts of AI systems adoption across demographic groups, e.g., age [20]. These findings highlight several gaps in the existing literature, warranting further future investigations in several areas.

D. Future Research Directions

The existing body of literature highlights significant gaps that necessitate future investigation in several areas. First, there is a need for cross-cultural trust dynamics studies, especially for conducting empirical studies across diverse cultural contexts. These studies are essential to comprehending how cultural values, norms, and communication styles interact with trust dynamics and how these cultural elements shape trust emergence, calibration, and repair in human-AI teams. Developing culturally adaptive AI interaction designs is another study area. Future research also needs to look at longitudinal studies. Literature highlights the importance of implementing long-term studies, especially in realistic operational settings, to observe how trust evolves. Also, to learn how to adapt to changing AI capabilities or task demands and recover (or fail to recover) from significant failures over time. In addition, current research underscores the importance of conducting a study to explore effective trust repair mechanisms. There is a need to move beyond simple apologies and explanations. It is time to investigate and develop robust trust repair strategies, potentially involving demonstrable AI learning and correction, adaptive transparency levels, or mechanisms for shared responsibility.

The next area that the thorough review reveals a compelling need to examine is calibration interventions. Literature highlights the need for evaluating and designing interfaces and training methods specifically aimed at improving trust calibration, helping users critically assess AI output, and avoiding over-reliance and under-reliance. These topics are significant in areas that require a cognitive load or time pressure. The development of diverse and effective AI error detection tools is also required. Their role should extend beyond identifying correct outputs to include error checking.

For the coming AI systems, researchers must further define and investigate the critical issues beyond task performance, including "teaming competencies". AI agents need to be perceived as effective collaborators by humans. Specifically, proactive communication, adaptivity, predictability, and expressing intent. Ethical frameworks for trust are more necessary than ever. Researchers must continue to explore and connect such frameworks to aspects of our lives in general and to the trustworthy cooperation of the operator AI in particular. The focus must address potential issues of manipulation, undue influence, and the balance between fostering trust through ethics and privacy, and security and between maintaining appropriate operator's vigilance and autonomy. The findings also underscore the importance of developing future AI models capable of tracking trust and addressing its dynamics, particularly within operator-AI team settings. These models must feature advanced capabilities that address bias, emotions, and social influence and have the ability to learn over time [72].

E. Solutions from Existing Literature

This review highlights significant gaps in the current understanding of trust dynamics in operator-AI collaboration, particularly regarding enhancing trust by increasing AI transparency and explainability (XAI), the need for adaptive trust calibration mechanisms, designing effective human-AI collaborations and teams, ensuring AI robustness, communicating uncertainty, and incorporating ethical AI principles and governance into design. The current literature offers emerging and promising directions for addressing these challenges. For example, the explorations of the need to develop human-centered XAI designs, presented in [80], can be adapted to provide insights into the problem of black-box systems, which are one of the reasons why decisions are difficult for operators to understand, undermining trust and complicating problem diagnosis. New designs provide understandable explanations (a posteriori and a priori), focusing on how these explanations are presented, such as through visual interfaces or natural language. The goal is to enhance operator understanding and trust. Similarly, implementing the mechanisms proposed by [81] can support adaptive trust calibration solutions for dynamic systems where operators rely on AI. These systems use real-time feedback about AI trust and uncertainty and targeted training interventions to help operators maintain appropriate trust levels. Similarly, to meet the requirements for effective human-AI teamwork and collaboration models, a potential solution may be [82]. It can provide a valuable perspective for reconsidering the weaknesses of human-AI collaboration caused by unclear roles, communication breakdowns, and a lack of shared understanding of tasks and constraints that can be addressed. According to this work, human-AI teamwork can be enhanced by focusing on joint task design where roles are optimally distributed based on strengths. This includes enabling shared mental models through AI design and implementing flexible autonomy for dynamic control transitions. Trust erosion is another gap caused by miscommunication or errors in AI. The work presented in [83], [84] can help address this trust erosion. AI systems are known to fail unexpectedly or behave unpredictably, significantly eroding user trust and potentially leading to risks with dire consequences. Integrating ethical AI principles and governance into all phases of the AI development lifecycle is the missing solution to this problem. Implementing bias detection techniques, fairness metrics, and privacy preservation will be helpful. Success in achieving this could lead to building systematic trust through responsible design and auditing, as proposed in [84], [85]. We still need to expand the scope of ethical principles. This expansion will address broader ethical concerns such as bias, fairness, and accountability. Addressing these concerns increases operator trust and facilitates AI adoption. These challenges and limitations pave the way for future research. Finally, Section VIII concludes with a review and summarizes the main arguments.

To summrize, scholarly articles in the review consistently highlighted essential and inspiring subjects. Handling these dreaming topics through continued interdisciplinary collaboration is the only way to bring them into reality. It is necessary to continue combining insights from computer science, human factors, psychology, sociology, and organizational sciences. Only in this way can we realize this bright future vision, one in which more reliable AI systems collaborate with a greater understanding of the dynamics of trust. Even more importantly, we can address imminent risks, such as those witnessed in HAL 9000 from the movie “2001: A Space Odyssey” story, which reminds us of the importance of calibrating and enhancing trust between humans and their artificial companions at all stages. While these gaps exist, the current literature offers emerging and promising directions for addressing these challenges.

VIII. Conclusions

Trust is dynamic, not static. It has a starting point and grows and wanes according to the influencing factors that intertwine to define its parameters. These significant factors include human perception, AI capabilities, and changing context. Given the importance of trust, we must know the best ways to maintain it. Indeed, we need to dig deeper to know more about the factors that lead to its rise and decline and the best strategies for repairing it if it is damaged. Despite significant progress, our current methods fall short of capturing the intricacies of trust, emphasizing the need for more work to explore the mysteries of trust in the new and rapidly expanding work environments. AI performance is often referred to as the foundation upon which trust in human-AI collaboration rests. Other equally essential elements include transparency, explainability, and accuracy. However, a closer look reveals that these foundations suffer from cognitive holes and require continued efforts to bridge existing gaps, several of which were revealed by this review. AI system transparency is an important and influential factor in collaborative work in human-AI teams. However, it was found that it is not always sufficient when we looked closely at the results of a case study on an XAI used as a Clinical Decision Support System to assess the importance and impact of transparency. In this study, operators (healthcare practitioners) were presented with explanations varying in clarity levels, ranging from start forward to semi-clear to inconsistent. The results demonstrated that simply providing explanations for transparency does not transform a black box problem into a white box and, most importantly, does not ensure proper trust calibration. In the study, some explanations reinforced overreliance on the system, while participants ignored others, especially unclear or inaccurate ones. These conclusions indeed raise our concern in such a high-stakes environment, and it is a wake-up call—people focused on providing information, not necessarily ensuring its compatibility with human understanding and critical evaluation. Digging deeper, it was found several factors influencing the nature of trust, some related to AI systems, others to operators and context. These multiple factors interact in a complex dynamic that ultimately enhances or degrades trust. Operators have inherent skill capabilities and influential cognitive limitations, including individual propensities, cognitive biases, and cultural backgrounds, whose role in shaping trust should not be overlooked. In addition, there are the required competencies for interacting with AI. While understanding the human factors involved in communication and teamwork is vital, this is insufficient www.ijacsa.thesai.org 12 — P a g e (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 16, No. 07, 2025 in a human-machine environment devoid of human emotions and social stimuli. On the contrary, this relationship with unstable and renewable factors presents new challenges, many of which still need to be explored. Why do the cross-cultural variations remain a glaring gray spot when partnering with AI? Similarly, are Western-centric models universal, or are studies overlooking crucial cultural nuances that drastically alter human-AI interactions? And how can people expect operators to develop calibrated trust when training often neglects to equip them with the critical skills for evaluating AI? Also, what would happen if we left operators without enough knowledge or when leaving them vulnerable to automation bias and misinterpretations? Many fundamental questions related to the human aspect need scientific answers. As highlighted in the literature, we need to transform operators from passive users into active, discerning partners who can interpret and measure AI trust and not just accept the system as a black box. To successfully measure trust across its various stages, different metrics will need to be used depending on the context and complexity, in addition to each factor that must be measured in both the human and machine elements. Our review highlights several important gaps that require further research. First, we urgently need new requirements for the responsible design and deployment of AI systems, especially given the rapid developments in this field. Furthermore, a practical methodology is needed to ensure that the design adheres to the new requirements and the principles of ethics, transparency, and accountability to achieve responsible and trustworthy AI. During the collaboration lifecycle, the dynamic nature of trust (its fragile emergence, difficult maintenance, rapid erosion, and arduous repair) further complicates matters. Current attempts at trust repair are still limited to apologies or basic explanations, which appear woefully inadequate. Can an apology or clear explanation mend trust damaged by a system perceived as unreliable or biased? Or do we need to explore more human-like elements of trust repair? Likewise, despite growing recognition that AI is moving beyond a simple tool to a teammate, existing theoretical models struggle to keep pace. Existing frameworks are often helpful but fail to capture the unique challenges of human-AI interdependence fully. Similarly, there is a need to bridge the widening gap toward developing future systems capable of comprehensively replacing humans or transforming their role to become more like a tool used in a specific context when needed. So far, the hope to accurately define and measure "trust" and "operator-AI collaboration". The current metrics still fail to capture the essence of this relationship. The limitations we face in this regard are not merely academic controversies; they are realities with dire consequences unless more is done to bridge the gaps. Miscalibrating the future of human-AI teams will lead to disastrous consequences. This threat is most acute in high-risk areas, including healthcare, transportation, and defense. Society is entering an era of risks cloaked in complacency and misunderstanding. Innovative solutions need to be prioritized over the promise of AI efficiency. The review adopted a systematic approach to analyzing the existing literature by collecting and categorizing research papers in the selected field (the dynamics of trust in the collaborative environment between operators and artificial intelligence) to reveal areas of research focus and, conversely, absence. This task was followed by reviewing and examining relevant theoretical foundations that addressed the relationship between humans and modern technologies, such as AI, within the framework of collaborative work. The methodology identified areas where findings conflicted or where different studies used incompatible definitions. Synthesizing the findings, it was possible to determine whether certain aspects of "trust" or "operator-AI collaboration" were consistently overlooked or only superficially addressed. The synthesized outcome allowed for a comprehensive mapping of the current landscape of human-AI trust research, revealing underexplored areas and methodological inconsistencies. The identified gaps provide a rationale for the need for a framework, and these findings directly shape the aspects that the framework this review aims to address. The review confirms that the current approach is insufficient and that we need a fundamental shift beyond the stereotypical view of trust as merely a given emotion. Instead, trust must become a valuable ranking acquired through careful design of future operators' behaviors, perceptions, and ethics. This recommendation is consistent with the review gaps and emphasizes the need for a new approach to address the human aspects of operator-AI partnerships effectively. Therefore, exploring an innovative framework to address these needs, informed by cognitive and affective factors, will go beyond an academic endeavor to a vital solution. Through it, the hope is to contribute to safeguarding the future of this rapidly transformative technology. The new framework promises to transcend current limitations and enhance trust in human-AI collaborations. The development of this framework will be the focus of my following methodological study.

References

[1] A. C. Clarke, *2001: a space odyssey*. Penguin, 2016.

[2] M. Jutel, M. Zemelka-Wiacek, M. Ordak, O. Pfaar, T. Eiwegger, M. Rechenmacher, and C. A. Akdis, "The artificial intelligence (ai) revolution: How important for scientific work and its reliable sharing." *Allergy*, vol. 78, no. 8, 2023.

[3] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ: Pearson, 2021.

[4] P. Grover, A. K. Kar, and Y. K. Dwivedi, "Understanding artificial intelligence adoption in operations management: insights from the review of academic literature and social media discussions," *Annals of Operations Research*, vol. 308, no. 1, pp. 177–213, 2022.

[5] F. Kitsios and M. Kamariotou, "Artificial intelligence and business strategy towards digital transformation: A research agenda," *Sustainability*, vol. 13, no. 4, p. 2025, 2021.

[6] S. Schmager, I. Pappas, and P. Vassilakopoulou, "Understanding human-centred ai: a review of its defining elements and a research agenda," *Behaviour & Information Technology*, 2024. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/0144929X.2024.2448719

[7] C. Dragomirescu-Gaina, D. Philippas, and M. G. Tsionas, "Trading off accuracy for speed: Hedge funds' decision-making under uncertainty," *International Review of Financial Analysis*, vol. 75, p. 101728, 2021.

[8] N. J. Mcneese, B. G. Schelble, L. Canonico, and M. Demir, "Who/what is my teammate? team composition considerations in human–ai teaming," *IEEE Transactions on Human-Machine Systems*, vol. 51, pp. 288–299, 2021.

[9] V. Hagemann, M. Rieth, A. Suresh, and F. Kirchner, "Human-ai teams—challenges for a team-centered ai at work," *Frontiers in Artificial Intelligence*, vol. 6, p. 1252897, 2023.

[10] H.-A. Teaming, "State-of-the-art and research needs," *National Academies of Sciences, Engineering and Medicine, Washington DC*, vol. 10, p. 26355, 2022.

[11] M. J. Sáenz, E. Revilla, and C. Simón, "Designing ai systems with human-machine teams," *MIT Sloan Management Review*, vol. 61, no. 3, pp. 1–5, spring 2020. [Online]. Available: https://www.proquest.com/scholarly-journals/designing-ai-systems-with-human-machine-teams/docview/2392464050/se-2

[12] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors: The Journal of Human Factors and Ergonomics Society, Human Factors, 46(1), 50–80. URL: https://doi.org/10.1518/hfes.46.1.50_30392*, 2004.

[13] The Interaction Design Foundation. (2025) What Is Human-Centered AI (HCAI)? Accessed: May 23, 2025. [Online]. Available: https://www.interaction-design.org/literature/topics/human-centered-ai

[14] A. Korinek and A. Balwit, "Aligned with Whom? Direct and Social Goals for AI Systems," National Bureau of Economic Research, Working Paper 30017, 2022. [Online]. Available: https://www.nber.org/system/files/working_papers/w30017/w30017.pdf

[15] O. Vereschak, G. Bailly, and B. Caramiaux, "How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–39, 2021.

[16] M. Naiseh, D. Al-Thani, N. Jiang, and R. Ali, "How the different explanation classes impact trust calibration: The case of clinical decision support systems," *Int. J. Hum. Comput. Stud.*, vol. 169, p. 102941, 2023.

[17] J. Reyes, A. U. Batmaz, and M. Kersten-Oertel, "Trusting ai: does uncertainty visualization affect decision-making?" *Frontiers in Computer Science*, vol. 7, p. 1464348, 2025.

[18] M. B. Junaid and M. B. Hassan, "Ai and human-robot interaction: A review of recent advances and challenges," *GSC Advanced Research and Reviews*, vol. 18, pp. 321–330, 2024. [Online]. Available: https://gsconlinepress.com/journals/gscarr/sites/default/files/GSCARR-2024-0070.pdf

[19] M. Y. M. Naser and S. Bhattacharya, "Empowering human-ai teams via intentional behavioral synchrony," *Frontiers in Neuroergonomics*, vol. 4, 2023.

[20] M. C. Dorton and S. A. Harper, "The trust calibration maturity model for characterizing and communicating trustworthiness of ai systems," *arXiv preprint arXiv:2503.15511*, 2025. [Online]. Available: https://arxiv.org/pdf/2503.15511

[21] D. D. Scholz, J. Kraus, and L. Miller, "Measuring the propensity to trust in automated technology: Examining similarities to dispositional trust in other humans and validation of the ptt-a scale," *International Journal of Human–Computer Interaction*, 2025. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10447318.2024.2307691

[22] A. Bostrom, J. L. Demuth, C. D. Wirz, M. G. Cains, A. Schumacher, D. Madlambayan, A. S. Bansal, A. Bearth, R. Chase, K. M. Crosman, D. J. Gagne II, S. L. Henderson, A. McGovern, R. J. Redmon, Y. Rao, and I. Ebert-Uphoff, "Trust and trustworthy artificial intelligence: A research agenda for ai in the environmental sciences," *Risk Analysis*, vol. 44, pp. 1364–1378, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/risa.14245

[23] S. E. Kase, R. S. G. III, W. A. III, and C. S. Morris, "Trust and trustworthiness in human-ai interaction: A review and proposed research agenda," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 66, no. 1, 2022, pp. 1473–1477.

[24] K. Okamura and S. Yamada, "Adaptive trust calibration for human-ai collaboration," *Plos one*, vol. 15, no. 2, p. e0229132, 2020.

[25] S. Benk, J. M. Becker, and A. Kluge, "Defining human-ai teaming the human-centered way: A scoping review and network analysis," *Frontiers in Artificial Intelligence*, vol. 6, p. 1250725, 2024. [Online]. Available: https://doi.org/10.3389/frai.2023.1250725

[26] S. Kang, A.-E. Potinteu, and N. Said, "Explainitai: When do we trust artificial intelligence? the influence of content and explainability in a cross-cultural comparison," *arXiv preprint arXiv:2503.17158*, 2025.

[27] C. Attig, P. Wollstadt, T. Schrills, T. Franke, and C. B. Wiebel-Herboth, "More than task performance: Developing new criteria for successful human-ai teaming using the cooperative card game hanabi," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–11.

[28] N. K. Dang, M. T. Spaan, and D. Lo, "How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. Article 426, 1–32, 2021. [Online]. Available: https://doi.org/10.1145/3479570

[29] C. Esterwood and L. P. Robert, "Repairing trust in robots?: A meta-analysis of hri trust repair studies with a no-repair condition," in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2025, pp. 410–419.

[30] H. Liu, V. Lai, and C. Tan, "Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, pp. 1 – 45, 2021.

[31] A. Laitinen and O. Sahlgren, "Ai systems and respect for human autonomy," *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[32] D. Dellermann, A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. Ebel, "The future of human-ai collaboration: a taxonomy of design knowledge for hybrid intelligence systems," *arXiv preprint arXiv:2105.03354*, 2021.

[33] E. E. Makarius, D. Mukherjee, J. D. Fox, and A. K. Fox, "Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization," *Organizations & Markets: Policies & Processes eJournal*, 2020.

[34] R. Bansal and A. Sangwan, "A comprehensive review of artificial intelligence," *International Journal For Multidisciplinary Research*, vol. 6, pp. 149–158, 2024. [Online]. Available: https://www.ijfmr.com/research-paper.php?id=28932

[35] M. H. Haag, S. R. Schmidthuber, D. F. Dietl, and M. R. R. Bitsch, "Exploring generative artificial intelligence: A taxonomy and types," in *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS)*. University of Hawaii at Manoa, 2024. [Online]. Available: https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/fa9a6175-9ff2-4ad4-868e-fec5127cd430/content

[36] M. M. Al Rashed, M. S. Uddin, A. K. M. L. Khan, P. K. Roy, M. M. Islam, and M. S. Hossain, "Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems," *Frontiers in Artificial Intelligence*, vol. 5, p. 803098, 2022.

[37] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Academy of management annals*, vol. 14, no. 2, pp. 627–660, 2020.

[38] E. Georganta and A.-S. Ulfert, "My colleague is an ai! trust differences between ai and human teammates," *Team Performance Management: An International Journal*, vol. 30, no. 1/2, pp. 23–37, 2024.

[39] J. Liu, X. Liu, D. Wang, and Y. Zhang, "Grading by AI makes me feel fairer? how different evaluators affect college students' perception of fairness," *Frontiers in Psychology*, vol. 15, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1221177/full

[40] A. G. Glikson and A. Kaplan, "Human trust in artificial intelligence: Review of empirical research," *Academy of Management Annals*, vol. 15, pp. 627–665, 2021. [Online]. Available: https://journals.aom.org/doi/full/10.5465/annals.2018.0057

[41] P. Puranam, "Human–ai collaborative decision-making as an organization design problem," *Journal of Organization Design*, vol. 10, pp. 75 – 80, 2020.

[42] R. Parasuraman, T. Sheridan, and C. Wickens, "A model for types and levels of human interaction with automation," *IEEE transactions on systems, man, and cybernetics. Part A, Systems and humans : a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 30 3, pp. 286–97, 2000.

[43] R. Parasuraman and D. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors: The Journal of Human Factors and Ergonomics Society*, vol. 52, pp. 381 – 410, 2010.

[44] F. M. Cau and L. D. Spano, "Exploring the impact of explainable ai and cognitive capabilities on users' decisions," *arXiv preprint arXiv:2505.01192*, 2025.

[45] V. Venkatesh and F. D. Davis, "A theoretical extension of the technology acceptance model: Four longitudinal field studies," *Management science*, vol. 46, no. 2, pp. 186–204, 2000.

[46] J. Dingel, A.-K. Kleine, J. Cecil, A. L. Sigl, E. Lermer, and S. Gaube, "Predictors of health care practitioners' intention to use ai-enabled clinical decision support systems: Meta-analysis based on the unified theory of acceptance and use of technology," *Journal of medical internet research*, vol. 26, p. e57224, 2024.

[47] M. Madsen and S. Gregor, "Measuring human-computer trust," vol. 53, pp. 6–8, 2000.

[48] E. Glikson and A. W. Woolley, "A systematic literature review of user trust in AI-enabled systems: An HCI perspective," *Human-Computer Interaction*, vol. 35, no. 10, pp. 933–1001, 2020. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/07370024.2020.1750577

[49] S. Tannenbaum and E. Salas, *Teams that work: The seven drivers of team effectiveness*. Oxford University Press, 2020.

[50] A. C. Costa, R. A. Roe, and T. Taillieu, "Trust within teams: The relation between individual trustworthiness, team trust, and team performance," *European Journal of Work and Organizational Psychology*, vol. 10, no. 3, pp. 225–244, 2001. [Online]. Available: https://doi.org/10.1080/13594320143000654

[51] M. R. Endsley, "Situation awareness misconceptions and misunderstandings," *Journal of Cognitive Engineering and Decision*

Making, vol. 9, no. 1, pp. 4–32, 2015. [Online]. Available: https://doi.org/10.1177/1555343415572631

[52] P. Esmaeilzadeh, "Use of ai-based tools for healthcare purposes: A survey study from consumers' perspectives," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–21, 2022. [Online]. Available: https://doi.org/10.1186/s12911-022-01891-8

[53] S. Kim, S. Choo, D. Park, H. Park, C. S. Nam, J.-Y. Jung, and S. Lee, "Designing an xai interface for bci experts: A contextual design for pragmatic explanation interface based on domain knowledge in a specific context," *International Journal of Human-Computer Studies*, vol. 174, p. 103009, 2023.

[54] J. G. G. De Visser, J. C. D. Van Der Meij, and M. W. B. Van Der Meij, "Inferring trust from users' behaviours: Agents' predictability positively affects trust, task performance and cognitive load in human-agent real-time collaboration," *Frontiers in Robotics and AI*, vol. 8, p. 642201, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2021.642201/full

[55] T. B. Sheridan, *Humans and automation: System design, responsibility, and allocation of function*, 2nd ed. John Wiley & Sons, 2019.

[56] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, and S. Gil-López and Daniel Molina and Randy Benjamins and Raja Chatila and Francisco Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020. [Online]. Available: https://doi.org/10.1016/j.inffus.2019.12.012

[57] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2017.

[58] S. Berretta, A. Tausch, G. Ontrup, B. Gilles, C. Peifer, and A. Kluge, "Defining human-ai teaming the human-centered way: a scoping review and network analysis," *Frontiers in Artificial Intelligence*, vol. 6, 2023.

[59] P. Malinowska and E. Alberdi, "Trust development and explainability: A longitudinal study with a personalized assistive system," *Algorithms*, vol. 8, no. 3, p. 20, 2024.

[60] D. Kishnani, "The uncanny valley: An empirical study on human perceptions of ai-generated text and images," Master's thesis, Massachusetts Institute of Technology, 2025. [Online]. Available: https://dspace.mit.edu/bitstream/handle/1721.1/159096/kishnani-deepalik-sm-sdm-2025-thesis.pdf?sequence=-1&isAllowed=y

[61] S. Gaube, H. Suresh, M. Raue, A. Merritt, S. J. Berkowitz, E. Lermer, P. Szolovits, and M. Ghassemi, "Do as ai say: susceptibility in deployment of clinical decision-aids," *npj Digital Medicine*, vol. 4, no. 1, p. 31, 2021. [Online]. Available: https://doi.org/10.1038/s41746-021-00385-9

[62] J. B. Lyons, K. Stokes, and M. D. McNeese, "Human trust in artificial intelligence: Review of empirical research," *Academy of Management Annals*, vol. 15, no. 2, pp. 627–652, 2021.

[63] S. Dhuliawala, V. Zouhar, M. El-Assady, and M. Sachan, "A diachronic perspective on user trust in ai under uncertainty," *arXiv preprint arXiv:2310.13544*, 2023.

[64] Z. Chen, Y. Luo, and M. Sra, "Higher stakes, healthier trust? an application-grounded approach to assessing healthy trust in high-stakes human-ai collaboration," *arXiv preprint arXiv:2503.03529*, 2025. [Online]. Available: https://arxiv.org/abs/2503.03529

[65] I. Gkikopouli and S. Kontogiannis, "A Critical AI View on Autonomous Vehicle Navigation: The Growing Danger," *Electronics*, vol. 13, no. 18, p. 3660, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/18/3660

[66] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International journal of cognitive ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.

[67] U.S. Department of Health & Human Services, "Trustworthy AI (TAI) Playbook: Executive Summary," 2021, accessed: [Insert Date Accessed, e.g., May 23, 2025]. [Online]. Available: https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook-executive-summary.pdf

[68] M. J. McGrath, A. Duenser, J. Lacey, and C. Paris, "Collaborative human-ai trust (chai-t): A process framework for active management of trust in human-ai collaboration," *arXiv preprint arXiv:2404.01615*, 2024. [Online]. Available: https://arxiv.org/abs/2404.01615

[69] J. Rehm, M. Söllner, and A. Kluge, "Defining human-ai teaming the human-centered way: a scoping review and network analysis," *Computers in Human Behavior Reports*, vol. 11, p. 100318, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10570436/

[70] L. Roeder, P. Hoyte, J. van der Meer, L. Fell, P. Johnston, G. Kerr, and P. Bruza, "A quantum model of trust calibration in human–ai interactions," *Entropy*, vol. 25, no. 9, p. 1362, 2023.

[71] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, 2015. [Online]. Available: https://www.nature.com/articles/nature14541

[72] D. Gopinath, T. Kulesza, L. Ma, J. S. Siegel, P. Stone, and J. Zinman, "Bayesian modeling of human–ai complementarity," *Proceedings of the National Academy of Sciences*, vol. 119, no. 10, p. e2111547119, 2022. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2111547119

[73] M. Cascella, J. Montomoli, V. Bellini, A. Vittori, H. Biancuzzi, F. Dal Mas, and E. G. Bignami, "Crossing the ai chasm in neurocritical care," *Computers*, vol. 12, no. 4, p. 83, 2023.

[74] S. Gao and Y. Yan, "Verbal or written? the impact of apology on the repair of trust: Based on competence- vs. integrity-based trust violation," *Frontiers in Psychology*, vol. 13, p. 884867, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.884867/full

[75] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000.

[76] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International Journal of Cognitive Ergonomics*, vol. 4, pp. 53–71, 2000. [Online]. Available: https://www.researchgate.net/publication/247502831_Foundations_for_an_Empirically_Determined_Scale_of_Trust_n_Automated_Systems

[77] M. Naiseh, D. Al-Thani, N. Jiang, and R. Ali, "How the different explanation classes impact trust calibration: The case of clinical decision support systems," *International Journal of Human-Computer Studies*, vol. 169, p. 102941, 2023.

[78] I. Popa, M.-M. Cioc, A. Breazu, and C. Popa, "Identifying sufficient and necessary competencies in the effective use of artificial intelligence technologies," *Amfiteatru Economic*, 2024.

[79] K. Chen, C. Gu, and J. Huang, "Calibrating trust in ai-assisted decision making," UC Berkeley School of Information, Capstone Project Report, 2020. [Online]. Available: https://www.ischool.berkeley.edu/sites/default/files/sproject_attachments/humanai_capstonereport-final.pdf

[80] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: Informing design practices for explainable ai user experiences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 2020, pp. 1–15.

[81] N. Srinivasan and E. Thomason, "Adjust for trust: Mitigating trust-induced inappropriate reliance on ai assistance," *arXiv preprint arXiv:2502.13321*, 2025.

[82] J.-H. Ju and S. Aral, "Collaborating with ai agents: Field experiments on teamwork, productivity, and performance," *Management Science (Forthcoming)*, 2025.

[83] H. Sun, Z. Li, J. Zhang, J. Fu, and J. Li, "Robustness in ai-generated detection: Enhancing resistance to adversarial attacks," *arXiv preprint arXiv:2505.03435*, 2025.

[84] M. Teufel, R. Pinsler, J. E. S. W. Reinders, A. Maier, and C. L. Buckley, "Improving counterfactual truthfulness for molecular property prediction through uncertainty quantification," *arXiv preprint arXiv:2504.02606*, 2025.

[85] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," *arXiv preprint arXiv:1906.01409*, 2021.