ReGeo ISSN: 2177-3246

TRUST IN THE ARTIFICIAL INTELLIGENCE SYSTEM VERSUS INTERPERSONAL TRUST MEDIATED BY ARTIFICIAL INTELLIGENCE: A SYSTEMATIC REVIEW OF CONCEPTUAL DISTINCTIONS IN HUMAN-AI COLLABORATION RESEARCH

CONFIANÇA NO SISTEMA DE INTELIGÊNCIA ARTIFICIAL VERSUS CONFIANÇA INTERPESSOAL MEDIADA PELA INTELIGÊNCIA ARTIFICIAL: UMA REVISÃO SISTEMÁTICA DAS DISTINÇÕES CONCEITUAIS NA PESQUISA DE COLABORAÇÃO HUMANO-IA

CONFIANZA EN EL SISTEMA DE INTELIGENCIA ARTIFICIAL VERSUS CONFIANZA INTERPERSONAL MEDIADA POR LA INTELIGENCIA ARTIFICIAL: UNA REVISIÓN SISTEMÁTICA DE LAS DISTINCIONES CONCEPTUALES EN LA INVESTIGACIÓN SOBRE COLABORACIÓN HUMANO-IA

https://doi.org/10.56238/revgeov17n5-153

Data de submissão: 16/04/2026 **Data de publicação:** 16/05/2026

Marcelo Caetano de Ribeiro e Melo[1], Marcos Alves da Rocha[2], Silvania Soares Brandão[3], Jordana Layara Alves Dias[4]

ABSTRACT

Human-AI collaboration research has grown substantially in scope and methodological diversity, yet the construct of trust is frequently operationalized without distinguishing between two conceptually distinct targets: trust directed at the artificial intelligence system itself, and interpersonal trust among human team members that is modulated by the AI's presence in the collaboration. This conflation undermines theoretical precision and limits the comparability of findings across studies conducted in different disciplines and design traditions. This study addresses three research questions: how empirical studies operationalize the target of trust in human-AI collaboration contexts; to what extent studies distinguish between trust directed at AI systems and interpersonal trust affected by AI presence; and what measurement instruments are deployed to capture these distinct trust targets. A PRISMA 2020-compliant systematic review was conducted using the Scopus database. An initial pool of 7,541 records was reduced through sequential domain filtering (6,592 excluded) and a trust-keyword prefilter (744 excluded), yielding 205 candidates. A supplementary rescan of the classifier bucket recovered 10 additional records, producing a total screened pool of 215. Full eligibility was assessed for 56 records; 10 studies were confirmed through full-text access and included in the synthesis. A three-category coding scheme was applied to normalize trust-target designations across all 10 included studies. Six studies (60%) direct trust at the AI system (Category 1); two studies (20%) direct trust at

[1] Ph.D. in Biofuel Science and Technology. Institution: Universidade Estadual de Montes Claros (UNIMONTES). E-mail: marcelo.melo@unimontes.br

[2] Master's degree in Economic Development and Business Strategy. Institution: Universidade Estadual de Montes Claros (UNIMONTES). E-mail: marcos.alves@unimontes.br

[3] Master's degree in Business Administration. Institution: Instituto Federal do Norte de Minas Gerais (IFNMG). E-mail: silvania.brandao@ifnmg.edu.br

[4] Specialist in Accounting Management and Corporate Controllership. Institution: Faculdades Integradas do Norte de Minas (FUNORTE). E-mail: jordana.dias@soufunorte.com.br

interpersonal relationships within human teams that are modulated by AI presence (Category 2); and two studies (20%) present an ambiguous or unspecified trust target (Category 3). Instrument reporting is markedly incomplete: only three studies provide a clearly named measurement procedure, four studies use non-psychometric or qualitative methods, and three studies show extraction gaps in which instruments likely exist but were not captured in the abstracted record. The dominant pattern in human-AI collaboration research is to treat trust as directed at the AI system, leaving the interpersonal dimension substantially underexplored. The three-category framework offered here functions as a reusable coding scheme for synthesizing trust research across heterogeneous designs.

Keywords: Human-AI Collaboration. Trust in Artificial Intelligence. Interpersonal Trust. Trust Target. Systematic Review. Trust Measurement.

RESUMO

A pesquisa sobre colaboração humano-IA cresceu substancialmente em escopo e diversidade metodológica, porém o construto de confiança é frequentemente operacionalizado sem distinguir entre dois alvo conceitualmente distintos: a confiança direcionada ao próprio sistema de inteligência artificial e a confiança interpessoal entre membros da equipe humana que é modulada pela presença da IA na colaboração. Essa confusão prejudica a precisão teórica e limita a comparabilidade dos achados em estudos realizados em diferentes disciplinas e tradições de design. Este estudo aborda três questões de pesquisa: como os estudos empíricos operacionalizam o alvo de confiança em contextos de colaboração humano-IA; em que medida os estudos distinguem entre confiança direcionada aos sistemas de IA e confiança interpessoal afetada pela presença da IA; e quais instrumentos de medição são utilizados para capturar esses alvos distintos de confiança. Uma revisão sistemática compatível com PRISMA 2020 foi conduzida usando a base de dados Scopus. Um pool inicial de 7.541 registros foi reduzido através de filtragem sequencial por domínio (6.592 excluídos) e um pré-filtro de palavras-chave de confiança (744 excluídos), resultando em 205 candidatos. Uma rescagem suplementar do balde de classificadores recuperou 10 registros adicionais, produzindo um pool total analisado de 215. A elegibilidade completa foi avaliada para 56 registros; 10 estudos foram confirmados por acesso ao texto completo e incluídos na síntese. Um esquema de codificação de três categorias foi aplicado para normalizar as designações de alvo de confiança em todos os 10 estudos incluídos. Seis estudos (60%) direcionam confiança ao sistema de IA (Categoria 1); dois estudos (20%) direcionam confiança aos relacionamentos interpessoais dentro de equipes humanas que são modulados pela presença da IA (Categoria 2); e dois estudos (20%) apresentam um alvo de confiança ambíguo ou não especificado (Categoria 3). O relato de instrumentos é notavelmente incompleto: apenas três estudos fornecem um procedimento de medição claramente nomeado, quatro estudos usam métodos não-psicométricos ou qualitativos, e três estudos apresentam lacunas de extração nas quais os instrumentos provavelmente existem, mas não foram capturados no registro abstraído. O padrão dominante na pesquisa de colaboração humano-IA é tratar confiança como direcionada ao sistema de IA, deixando a dimensão interpessoal substancialmente inexplorada. O marco de três categorias oferecido aqui funciona como um esquema de codificação reutilizável para sintetizar pesquisas sobre confiança em designs heterogêneos.

Palavras-chave: Colaboração Humano-IA. Confiança em Inteligência Artificial. Confiança Interpessoal. Alvo de Confiança. Revisão Sistemática. Medição de Confiança.

RESUMEN
La investigación sobre colaboración humano-IA ha crecido sustancialmente en alcance y diversidad metodológica, sin embargo, el constructo de confianza frecuentemente se operacionaliza sin distinguir entre dos objetivos conceptualmente distintos: la confianza dirigida al propio sistema de inteligencia artificial y la confianza interpersonal entre miembros del equipo humano que es modulada por la presencia de la IA en la colaboración. Esta confusión socava la precisión teórica y limita la comparabilidad de los hallazgos entre estudios realizados en diferentes disciplinas y tradiciones de diseño. Este estudio aborda tres preguntas de investigación: cómo los estudios empíricos operacionalizan el objetivo de confianza en contextos de colaboración humano-IA; en qué medida los estudios distinguen entre confianza dirigida a sistemas de IA y confianza interpersonal afectada por la presencia de la IA; y qué instrumentos de medición se utilizan para capturar estos objetivos distintos de confianza. Se realizó una revisión sistemática compatible con PRISMA 2020 utilizando la base de datos Scopus. Un conjunto inicial de 7.541 registros se redujo a través del filtrado secuencial por dominio (6.592 excluidos) y un prefiltro de palabras clave de confianza (744 excluidos), lo que resultó en 205 candidatos. Un nuevo escaneo suplementario del depósito del clasificador recuperó 10 registros adicionales, produciendo un conjunto total examinado de 215. Se evaluó la elegibilidad completa para 56 registros; 10 estudios fueron confirmados mediante acceso al texto completo e incluidos en la síntesis. Se aplicó un esquema de codificación de tres categorías para normalizar las designaciones del objetivo de confianza en los 10 estudios incluidos. Seis estudios (60%) dirigen la confianza al sistema de IA (Categoría 1); dos estudios (20%) dirigen la confianza a las relaciones interpersonales dentro de equipos humanos que son modulados por la presencia de la IA (Categoría 2); y dos estudios (20%) presentan un objetivo de confianza ambiguo o no especificado (Categoría 3). La presentación de instrumentos es notablemente incompleta: solo tres estudios proporcionan un procedimiento de medición claramente nombrado, cuatro estudios utilizan métodos no psicométricos o cualitativos, y tres estudios muestran brechas de extracción en las que probablemente existan instrumentos pero no se capturaron en el registro extraído. El patrón dominante en la investigación de colaboración humano-IA es tratar la confianza como dirigida al sistema de IA, dejando la dimensión interpersonal sustancialmente inexplorada. El marco de tres categorías ofrecido aquí funciona como un esquema de codificación reutilizable para sintetizar investigaciones sobre confianza en diseños heterogéneos.

Palabras clave: Colaboración Humano-IA. Confianza en Inteligencia Artificial. Confianza Interpersonal. Objetivo de Confianza. Revisión Sistemática. Medición de Confianza.

1 INTRODUCTION

The integration of artificial intelligence into collaborative work settings has transformed the organizational conditions under which teams accomplish knowledge-intensive tasks. From AI-assisted decision support in healthcare and autonomous agents embedded in logistics operations, to generative AI tools that participate directly in creative and project management workflows, AI systems are increasingly positioned not as peripheral tools but as active participants in task-oriented collaboration (SEEBER *et al.*, 2020; MCNEESE *et al.*, 2018). The practical consequence of this transition is that the study of human-AI collaboration is no longer adequately served by models developed for human-computer interaction in dyadic, tool-use contexts. Research on teams that include AI members requires frameworks capable of addressing the social, cognitive, and organizational dimensions that emerge when AI agents assume teammate roles (DELLERMANN *et al.*, 2019; O'NEILL *et al.*, 2022).

Within this transformed landscape, trust has come to occupy a central explanatory position. Decades of research in human factors and organizational psychology have established that trust governs whether human operators accept automated system outputs, integrate AI recommendations into consequential decisions, and maintain reliance relationships calibrated to actual system capabilities (LEE; SEE, 2004). Meta-analytic evidence confirms that trust is one of the most consistently examined constructs in the human-automation and human-autonomy teaming literatures, serving as a mediator of performance outcomes across a wide range of task types, system modalities, and operator characteristics (HANCOCK *et al.*, 2011). More recently, as generative AI has entered professional workflows, studies have examined how trust in AI tools shapes employee innovation, organizational learning, and team dynamics, extending the trust construct into organizational behavior and management science (LIN; WANG; SHENG, 2025; ZHOU; LYU, 2026).

Yet beneath this apparent conceptual coherence, the literature conceals a significant structural ambiguity. The term “trust” in human-AI collaboration research is routinely used without specifying its primary target, that is, the entity or relationship toward which trust is directed. This omission matters because the psychological, social, and organizational processes involved in trusting an AI system are fundamentally different from those involved in trusting a human colleague who happens to operate alongside an AI partner. Trust in an AI system concerns the calibration of expectations about system competence, reliability, and behavioral consistency in the face of uncertainty (LEE; SEE, 2004). Trust between human

team members that is modulated by AI presence in the team involves interpersonal processes including perceived similarity, social identification, affective bonds, and the relational consequences of having an AI member occupy a social role that was previously held by a human (GEORGANTA; ULFERT, 2024; MAYER; DAVIS; SCHOORMAN, 1995).

When studies conflate these two distinct trust targets, the consequences are both theoretical and practical. Theoretically, comparisons across studies become unreliable because studies ostensibly examining “trust in human-AI collaboration” may be measuring qualitatively different constructs. Interventions designed to build trust of one type may inadvertently reduce trust of the other type, and theoretical frameworks requiring distinct treatment of each trust form are incorrectly applied. Practically, organizational assessments of AI adoption that rely on surveys of trust in AI technology capture only one dimension of what is happening to trust in teams that include AI members; the interpersonal layer remains invisible to those instruments.

Despite the growing body of research on human-AI collaboration, no systematic effort has yet mapped how empirical studies operationalize the trust-target construct across the literature. Reviews of trust in automation (HANCOCK *et al.*, 2011; LEE; SEE, 2004) have addressed general dimensions of trust in technology but were developed before the emergence of generative AI and team-embedded AI contexts. Conceptual work addressing the gap between theoretical trust definitions and practical reliance behavior (NIEMANN; SCHMID, 2025) does not provide a coding framework applicable to existing studies at scale.

This paper addresses the gap through a PRISMA 2020-compliant systematic review of empirical studies on trust in human-AI collaboration, applying a three-category normalization scheme to each study's trust-target operationalization. The three categories are: Category 1 (trust in AI), referring to trust directed at the AI system or agent as its primary object; Category 2 (interpersonal trust mediated by AI), referring to trust between human team members that is modulated or structured by the AI's presence as a collaborating agent; and Category 3 (ambiguous or unspecified), applied when a trust construct is present but the target cannot be unambiguously assigned to either of the first two categories.

Three research questions guide the review:

RQ1: How do empirical studies operationalize the target of trust in human-AI collaboration contexts?

RQ2: To what extent do studies distinguish between trust directed at AI systems and interpersonal trust affected by AI presence?

RQ3: What measurement instruments are deployed to capture these distinct trust targets across the human-AI collaboration literature?

The three research questions are addressed through a systematic search of the Scopus database, a PRISMA 2020 screening procedure, and a two-step normalization and reconciliation analysis applied to the 10 full-text-confirmed studies that met the eligibility criteria. The remainder of the paper is organized as follows: Section 2 presents the theoretical background, including foundational trust models and the conceptual basis for the three-category coding scheme; Section 3 describes the methodology; Section 4 reports findings and discusses their implications; and Section 5 presents the conclusion.

2 THEORETICAL BACKGROUND

2.1 TRUST IN AUTOMATION: FOUNDATIONAL MODELS

The systematic study of trust in automated systems has a well-established theoretical lineage in human factors engineering, cognitive engineering, and organizational psychology. Lee and See (2004, p. 51) define trust in automation as “the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability.” Three structural features of this definition have particular bearing on the present review. First, trust is an attitude rather than a behavior, meaning it is a psychological disposition that mediates behavioral outcomes (such as reliance or override decisions) rather than constituting such outcomes directly. Second, trust is inherently relational: it is directed at a specific agent, implying that the trustor-trustee relationship must be specified for the construct to be well-defined. Third, trust arises under conditions of uncertainty and vulnerability, which means that measurement designs that eliminate uncertainty (for example, through high-fidelity simulations with known performance parameters) may depress or artificially constrain the trust construct they seek to assess.

Lee and See (2004) articulate a calibration model of trust as the primary normative standard for human-automation interaction. Appropriate reliance on automation requires not merely high trust but trust that is calibrated to the actual capabilities and limitations of the system. Overtrust produces misuse, the uncritical acceptance of faulty or inappropriate AI outputs that a properly calibrated human would override. Undertrust produces disuse, the rejection of AI capabilities that would genuinely benefit the human-AI collaboration. Both forms of miscalibration reduce collaboration effectiveness, and both can persist when users lack feedback about system performance across tasks or when system reliability is

inconsistent across contexts (NAISEH *et al.*, 2021). The calibration framing establishes that trust in automation is a dynamic, ongoing process rather than a stable trait, making its measurement particularly sensitive to design choices regarding timing, task context, and feedback availability.

Mayer, Davis and Schoorman (1995) provide the most influential general model of organizational trust, identifying three antecedents of willingness to be vulnerable to another party: ability (the trusted party's competence in the domain relevant to the trusting relationship), benevolence (the extent to which the trusted party is motivated to act in the trustor's interests rather than merely in its own), and integrity (the degree to which the trusted party adheres to principles the trustor finds acceptable). This framework was originally developed for interpersonal trust in organizational hierarchies, but has been extensively adapted to human-AI settings. In AI contexts, ability corresponds to the system's technical competence and task performance quality; benevolence is interpreted as the alignment of the AI system's optimization targets with the user's interests and values; and integrity approximates AI system predictability, transparency, and adherence to user-defined operating parameters (NAISEH *et al.*, 2021; NIEMANN; SCHMID, 2025). A key tension in applying the Mayer, Davis and Schoorman (1995) model to AI systems concerns the benevolence dimension: traditional AI systems lack intentionality and are not “motivated” in any phenomenological sense, which complicates the direct attribution of benevolence and requires researchers to specify what organizational or design proxies they use to operationalize this dimension in AI contexts.

Hancock and colleagues (2011) substantially advance the empirical foundation of trust in automation research through a meta-analysis of factors affecting trust in human-robot interaction, synthesizing findings across 49 studies. Their analysis organizes antecedents of trust into three clusters: robot-related factors (including performance, presentation, and design attributes), environment-related factors (task type, complexity, and operational context), and human-related factors (prior experience, individual differences in disposition to trust, and demographic characteristics). A major finding is that robot performance factors receive substantially more research attention than human-related factors, despite meta-analytic evidence that human-related factors have comparable effect magnitudes. This methodological asymmetry is relevant to the present review because it suggests the field has historically concentrated on the system side of the trust relationship, potentially

underrepresenting the contribution of human psychological and social factors, including those involved in interpersonal trust modulation.

2.2 TRUST CALIBRATION, EXPLAINABILITY, AND DESIGN

A central concern in contemporary human-AI collaboration research is the relationship between AI explainability and trust calibration. The normative assumption that explaining AI reasoning to human users will improve trust calibration has generated a substantial body of experimental research, with mixed results. Naiseh and colleagues (2021), in a qualitative study using think-aloud protocols with healthcare professionals interacting with an AI recommendation system, find that explainable AI systems do not automatically produce well-calibrated trust. Users exhibit systematic trust calibration errors, including skipping available explanations entirely and misapplying explanations to decision contexts for which they were not designed. The study identifies five design principles for XAI systems intended to support appropriate trust calibration: designing for genuine user engagement with explanations, challenging habitual decision patterns that bypass explanations, using attentional guidance mechanisms, introducing deliberate friction to slow automated acceptance, and embedding support for user training and learning about system behavior over time.

This line of research is directly relevant to the present review because it locates trust calibration at the interface between cognitive evaluation of system behavior and the design affordances that shape that evaluation. Critically, all of the trust constructs examined in the Naiseh and colleagues (2021) framework are Category 1 constructs: the trust object is the AI system's recommendations, and the miscalibration at issue concerns the human-AI dyad rather than any interpersonal dimension. The theoretical richness of the calibration framework does not, in itself, address what happens to interpersonal trust within the team when an AI system's recommendations are introduced or when its explanations are accepted or rejected.

Niemann and Schmid (2025) address a related but distinct conceptual problem in their review of trust frameworks in AI-assisted code generation: the theoretical and operational gap between trust (a psychological attitude, assessed through self-report or behavioral proxies) and reliance (an observable behavioral outcome, such as adopting versus overriding AI output). They argue that much of the apparent inconsistency in the trust-in-AI literature arises from conflating these two constructs, with some studies using behavioral reliance as a proxy for attitudinal trust. Their contribution is to provide a conceptual clarification that helps

reconcile apparently contradictory findings rather than generating new primary empirical evidence.

2.3 HUMAN-AI COLLABORATION FRAMEWORKS

Human-AI collaboration has been conceptualized through several complementary theoretical frameworks that vary in their emphasis on team structure, cognitive role distribution, and the degree of AI agency and social presence.

Seeber and colleagues (2020) present a research agenda for AI as a teammate, distinguishing conceptually between AI-as-tool (AI serving as an instrument under complete human control), AI-as-member (AI with delegated task responsibilities and partial autonomy), and AI-as-teammate (AI with shared goals, social presence, and dynamic role adaptation). This progression is not merely technological but social: as AI moves along the tool-to-teammate continuum, the relevant trust construct shifts from trust in a capability to trust in a social actor, bringing interpersonal trust dynamics into the collaboration space. Seeber and colleagues (2020) explicitly anticipate the need for research designs that can simultaneously assess trust in the AI member and trust among the human members who share the team with that AI member.

McNeese and colleagues (2018) examine human-autonomy teaming (HAT) with specific attention to how team mental models, communication patterns, and trust structures evolve in teams that include an autonomous synthetic teammate. Their experimental study finds that teaming with a synthetic agent modifies the interpersonal dynamics of the human members in ways that are not fully captured by traditional human-automation interaction models. McNeese and colleagues (2018) observe that trust in the autonomous agent can be operationally distinct from trust among the human team members, and that both forms of trust are simultaneously active in HAT contexts. Their work is foundational for the Category 2 construct in the present review: it establishes empirically that AI's presence in a team can reshape the interpersonal trust landscape among the human members, not only the individual reliance decisions of each human vis-à-vis the AI.

Dellermann and colleagues (2019) develop the hybrid intelligence framework, which positions the combination of human and AI capabilities as producing collective intelligence exceeding what either party achieves independently. In the hybrid intelligence architecture, trust relationships are inherently multi-directional: human team members must develop calibrated trust in AI recommendations; the AI system must be designed to appropriately

weight and defer to human judgment in domains where human expertise exceeds AI capability; and human team members must maintain effective interpersonal trust relationships to coordinate their complementary contributions. Dellermann and colleagues' (2019) framework thus implies that Category 1 and Category 2 trust are not independent but are coupled in complex ways, with miscalibration in one dimension potentially cascading into the other.

O'Neill and colleagues (2022) provide the most comprehensive empirical review of human-autonomy teaming research to date, covering 128 studies published over more than a decade. Their analysis identifies trust as one of the most consistently examined constructs in the HAT literature, appearing in over half of all studies reviewed. Critically, they note significant heterogeneity in how trust is conceptualized and operationalized, and specifically flag the need for researchers to clarify whether the trust constructs studied concern the human-agent relationship (equivalent to Category 1), intra-team human relationships (equivalent to Category 2), or the team-level collective as a sociotechnical unit. The O'Neill and colleagues (2022) review thus provides direct scholarly precedent for the three-category coding scheme introduced in the present work.

2.4 THE THREE-CATEGORY FRAMEWORK: CONCEPTUAL FOUNDATIONS AND OPERATIONAL CRITERIA

The conceptual gap documented in the preceding sections motivates the three-category coding scheme that is the primary methodological contribution of this review. The scheme is designed to be operational rather than merely taxonomic: each category is defined by observable coding criteria applicable to published studies, making the scheme reproducible and auditable.

Category 1: Trust in AI. The primary object of trust is the AI system, agent, or algorithm. The construct captures the trustor's willingness to rely on the AI's outputs, recommendations, or actions in the face of uncertainty. Trust is operationalized as a dyadic relationship between a human user, operator, or team member and the AI system. Representative trust constructs in this category include trust in algorithmic decision-making, calibrated trust in AI recommendations, trust in autonomous agents, and AI-directed reliance. The psychological antecedents of Category 1 trust are the ability, benevolence, and integrity dimensions of the AI system as perceived by the human trustor (MAYER; DAVIS; SCHOORMAN, 1995), the system's perceived reliability and predictability (LEE; SEE, 2004),

and individual difference factors in disposition to trust technology (HANCOCK *et al.*, 2011). Category 1 trust studies appropriately employ psychometric instruments calibrated to human-system interaction contexts, behavioral measures of automation acceptance, and physiological or neuroimaging correlates of trust-related cognitive states.

Category 2: Interpersonal Trust Mediated by AI. The primary object of trust is the relationship between human team members. The AI's presence in the team functions as an independent variable, moderator, mediator, or contextual factor that modifies this interpersonal trust relationship. The dependent variable in a Category 2 study is the level or quality of trust between humans in the team, not trust in the AI itself. This category aligns with the social and organizational trust literature (MAYER; DAVIS; SCHOORMAN, 1995) and with human-autonomy teaming research that foregrounds team-level processes (MCNEESE *et al.*, 2018; SEEBER *et al.*, 2020). Category 2 trust studies appropriately employ instruments developed for interpersonal trust in team settings, such as multidimensional scales capturing affective and cognitive components of intra-team trust, and experimental designs in which team composition (with or without an AI member) is the manipulated variable.

Category 3: Ambiguous or Unspecified. A trust construct is present in the study, but the target cannot be unambiguously assigned to Category 1 or Category 2 based on the study's reported operationalization. This category encompasses three sub-types. First, studies that treat trust as a systemic property of a sociotechnical arrangement without decomposing it into directional components (trust in the system as a whole rather than in any specific agent within it). Second, studies that use trust-adjacent constructs (for example, suspicion, reliance, or confidence) that do not map cleanly onto the directed trust relationship implied by Categories 1 and 2. Third, qualitative studies in which trust is discussed thematically without a clear directional operationalization, such that the coding decision would depend on interpretive choices beyond what the study itself specifies.

Category 3 is not a residual or fail category; it is a substantively meaningful classification that signals theoretical underdetermination in the study's trust construct. A high proportion of Category 3 studies in a corpus would indicate that the field has not yet achieved sufficient conceptual clarity to specify trust targets in a replicable way, which is itself an important finding.

The distinction between Category 1 and Category 2 is not merely taxonomic. It carries direct implications for research design, measurement, and intervention. A finding that “trust is positively associated with team performance” has opposite policy implications depending on

whether trust is Category 1 (a finding about automation acceptance that suggests investing in AI transparency and reliability) or Category 2 (a finding about team cohesion that suggests investing in onboarding processes for AI-augmented teams and managing the social disruption of AI-as-teammate). Applying a Category 1 instrument to a Category 2 research question constitutes a construct validity threat; applying a Category 2 instrument to a Category 1 study may artificially suppress the variance attributable to the AI trust relationship by mixing it with interpersonal variance.

2.5 THE MEASUREMENT CHALLENGE IN HUMAN-AI TRUST RESEARCH

The measurement of trust in human-AI contexts presents challenges that distinguish it from trust measurement in purely interpersonal or purely machine-interaction contexts. Lee and See (2004) identify two broad classes of trust measures: self-report psychometric instruments and behavioral proxies. Self-report instruments range from validated multi-item scales (such as the Trust in Automation scale developed by Merritt and Ilgen; the Interpersonal Trust at Work scale; and domain-specific adaptations of the Mayer, Davis and Schoorman model) to ad hoc single-item measures constructed for specific experimental scenarios. Behavioral proxies include automation acceptance rates (the proportion of AI recommendations the human accepts versus overrides), response time distributions that reveal reliance tendencies, co-monitoring behaviors, and physiological indicators such as heart rate variability and EEG-derived indices that correlate with trust-related cognitive states (BALES *et al.*, 2025).

Each measurement approach carries distinct validity assumptions. Self-report measures are sensitive to social desirability, susceptibility to demand characteristics in experimental settings, and the limitation that participants may not have direct introspective access to the attitudes being assessed. Behavioral measures avoid these limitations but require careful design to ensure that the behavior observed (for example, following an AI recommendation) actually reflects the trust attitude rather than other causal factors such as time pressure, fatigue, or the perceived cost of verification. The convergence between self-report and behavioral trust measures is imperfect, and the conditions under which the two measures diverge are themselves theoretically informative (NIEMANN; SCHMID, 2025).

In AI-augmented teams, the measurement problem is compounded by the need to assess both Category 1 and Category 2 trust simultaneously within the same experimental or field setting, while maintaining independence between the two measurement channels. A

human team member's trust in an AI teammate may partly determine their interpersonal trust in their human teammates, and vice versa. Standard research designs that measure only one of the two trust dimensions without controlling for the other may produce biased estimates of both. This bidirectional confound is one of the structural reasons for the dominance of Category 1 designs in the literature: designs that treat only Category 1 trust are simpler to execute and less prone to construct confound than designs requiring simultaneous treatment of both trust targets.

This measurement challenge motivates the third research question (RQ3) and the instrument reconciliation procedure (FR2) applied in the present review.

3 METHODOLOGY

3.1 REVIEW DESIGN AND REPORTING STANDARD

This study follows the PRISMA 2020 guidelines for reporting systematic reviews (PAGE *et al.*, 2021). PRISMA 2020 was selected over its predecessor versions because it explicitly provides for searches conducted in non-database sources, supplementary search rounds, and classifier audit procedures, all of which feature in the present review's search architecture. The review addresses three research questions (RQ1, RQ2, RQ3) defined in Section 1. The unit of analysis is the individual empirical or conceptual study that includes a trust construct operationalized in the context of human-AI collaboration.

The review was conducted without restriction to experimental or quantitative designs. This methodological pluralism was deliberate: the research questions concern how the trust-target construct is operationalized across the diversity of existing work, and restricting to experimental designs would systematically exclude qualitative studies that may nevertheless contain important Category 1, 2, or 3 trust constructs. The cost of inclusiveness is a heterogeneous corpus that does not support formal meta-analytic synthesis; the benefit is a more complete picture of how trust-target operationalization varies across design types.

3.2 ELIGIBILITY CRITERIA

Studies were included if they met all five of the following criteria:

IC1 (Human-AI collaboration context): The study addresses a context in which one or more human actors and one or more AI systems interact in a task-oriented arrangement. This includes contexts described as human-autonomy teaming, AI-assisted decision-making, AI-augmented team work, AI-enabled project management, human-robot interaction in

collaborative tasks, and analogous formulations. Pure AI systems research without a human collaborator component was excluded.

IC2 (Trust construct present): The study includes a trust construct that is operationalized as a dependent variable, independent variable, moderator, mediator, or central theoretical element. Studies in which trust appears only incidentally in the introduction or as a background citation were not considered to meet this criterion.

IC3 (Scopus indexing): The study is indexed in the Scopus database at the time of the search.

IC4 (English language): The study is published in English.

IC5 (Evidence verifiability): Evidence supporting inclusion could be verified through full-text access, abstract-level assessment, or secondary source documentation.

Studies were excluded under two exclusion criteria:

EC1 (Deception design): The study employs a deception-based experimental design in which trust is formed under conditions of deliberate information manipulation that would not generalize to naturalistic trust formation processes.

EC2 (Trust absent or peripheral): The trust construct is absent from or entirely peripheral to the study's stated objectives and substantive contribution.

3.3 SEARCH STRATEGY

The search was conducted in the Scopus database using a structured query combining terms from three semantic clusters: (i) human-AI collaboration and teaming (including human-autonomy teaming, AI-augmented teams, AI-embedded workflows, and synthetic teammate constructs); (ii) trust, reliability, confidence, and reliance in AI and automated systems; and (iii) organizational, task performance, and team outcome terms relevant to human-AI collaboration effectiveness.

3.4 SCREENING AND SELECTION

The initial Scopus search yielded 7,541 records. Screening proceeded through a multistage sequential protocol.

In the first stage, domain filtering was applied to remove records outside the substantive scope of human-AI collaboration and team-based AI use. This stage excluded 6,592 records, retaining 949 records for further consideration.

In the second stage, a trust-keyword prefilter (T4) retained only records containing terms associated with trust, reliability, or confidence in human-AI interaction contexts. This prefilter excluded 744 records, yielding 205 records for the main screening phase.

Recognizing that keyword-based prefiltering may inadvertently exclude records in which trust is a secondary but relevant construct, a supplementary audit was conducted on the classifier bucket of 1,117 records that had passed the domain filter but been excluded at the T4 stage or in adjacent screening rounds. This supplementary rescan evaluated each record against IC1 and IC2, recovering 10 additional records that met both criteria despite the absence of explicit trust keywords in their indexed metadata. These 10 records were added to the main screening pool, producing a combined total of 215 records submitted to full screening.

Screening at the title and abstract level excluded 159 records for failure to meet the eligibility criteria, leaving 56 records for full eligibility assessment. Of the 56 records assessed for eligibility, one was excluded under EC1 due to a deception-based experimental design structurally incompatible with naturalistic trust formation. The 55 remaining eligible records were differentiated by evidence tier: 10 records were confirmed through direct full-text access; 38 were assessed at the abstract level only, with full text unavailable; and 7 were assessed through secondary source documentation. Consistent with the review's evidentiary standards, only the 10 full-text-confirmed studies were included in the final synthesis. Figure 1 presents the PRISMA 2020 flow diagram.

3.5 TRUST-TARGET NORMALIZATION PROCEDURE (FR1)

A systematic coding procedure was developed to normalize the trust-target field across all 10 included studies into the three-category scheme defined in Section 2.4. For each study, the primary constructs labeled as trust in the study's methods section, operationalization statement, and reported findings were evaluated against three sequential criteria:

Criterion A (AI-directed trust): Is the study's operationalization of trust structured as a dyadic relationship between a human actor and the AI system, with the AI system as the direct referent of the trust attitude? If yes, assign Category 1.

Criterion B (Interpersonal trust modulated by AI): Is the study's primary trust outcome a relationship between human team members, with the AI's presence serving as a structural or contextual variable (independent variable, moderator, or experimental condition) that shapes the interpersonal trust dynamic? If yes, assign Category 2.

Criterion C (Indeterminate): If neither Criterion A nor Criterion B can be applied unambiguously based on the reported operationalization, or if the study's trust construct is explicitly systemic or relational without specifying a directional human-to-agent target, assign Category 3.

One study required reclassification from its raw coded value in the extraction database. Georganta and Ulfert (2024) had been assigned a Category 1 value in the extraction record, reflecting a surface reading of the study as addressing AI-directed trust. Full-text review revealed a fundamentally different research design: both pre-registered experiments (total n = 828) used team composition (human-AI versus human-human) as the independent variable and measured interpersonal trust between human team members (operationalized through affective and cognitive trust subscales) as the dependent variable. The AI's presence was the experimental manipulation, not the trust object. The study was accordingly reclassified to Category 2. This reclassification, its evidentiary basis, and its implications for the frequency distribution are documented and disclosed in Section 4.5.

3.6 INSTRUMENT RECONCILIATION PROCEDURE (FR2)

The eligibility criteria required that each included study present a named or otherwise identifiable trust measurement instrument as a condition of passage through the screening process. Upon systematic review of the 10 included studies, a discrepancy was identified between the eligibility filter requirement and the actual instrument documentation available in the abstracted extraction record. A reconciliation procedure was applied to each study, distinguishing three conditions:

Condition I (Instrument documented): The study's extracted record includes a clearly named psychometric instrument, a described behavioral measurement procedure, or a physiological measurement protocol explicitly linked to trust assessment.

Condition II (Instrument inferred): The study's design characteristics (for example, a pre-registered experiment or a large-sample validated survey) make instrument presence highly probable, but the specific instrument name was not captured in the abstracted record. This condition is classified as extraction incompleteness rather than genuine instrument absence.

Condition III (No psychometric instrument applicable): The study's design is qualitative, conceptual, or relies on a trust-adjacent instrument not designed as a primary trust scale.

Instrument reconciliation results are reported in Table 3 of Section 4.

3.7 DATA ANALYSIS APPROACH

Following normalization and instrument reconciliation, frequency distributions were computed for trust-target categories (RQ1, RQ2) and for instrument documentation conditions (RQ3). Study characteristics including research design, sample size, country of origin, sector, and primary trust construct were tabulated in the evidence matrix (Table 1). No formal statistical inference tests were performed; the analytic approach is descriptive and interpretive, consistent with the qualitative nature of the coding scheme and the size of the corpus. Synthesis proceeds through a combination of conceptual mapping (grouping studies by trust-target category) and within-category comparative analysis (identifying similarities and contrasts in operationalization and measurement across studies assigned to the same category).

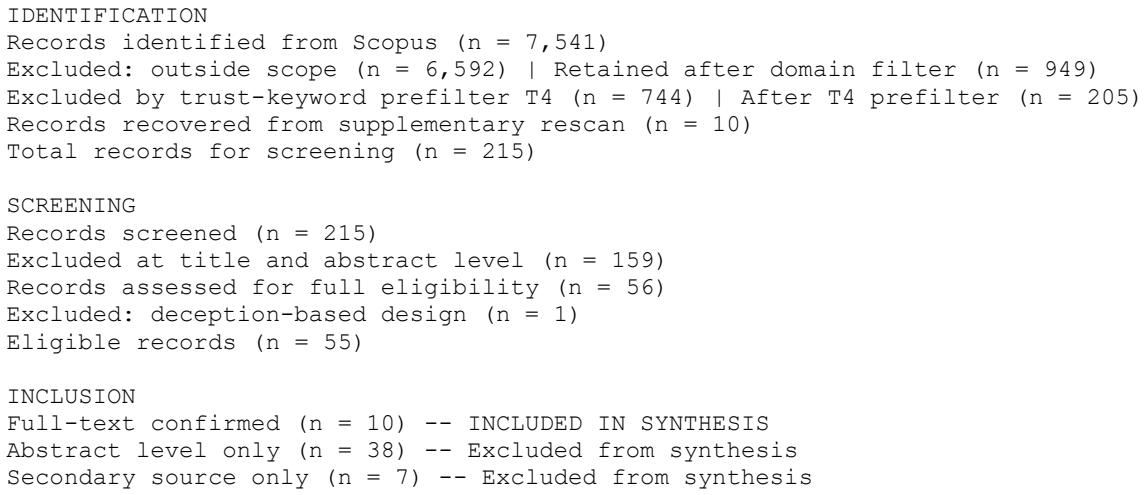
4 RESULTS AND DISCUSSION

4.1 PRISMA FLOW

Figure 1

Presents the complete PRISMA 2020 flow diagram for this review.

Figure 1. PRISMA 2020 flow diagram



Source: Prepared by the author following Page *et al.* (2021).

The funnel structure of the search protocol deserves comment. The two-stage filtering (domain filter followed by trust keyword prefilter) reduced the initial pool of 7,541 records by 90.8%, from 7,541 to 705 records, before the supplementary rescan. This high attrition

reflects the specificity of the intersection between human-AI collaboration and trust constructs: the broader literature on human-AI collaboration is substantially concerned with performance optimization, interface design, cognitive load, and system architecture, with trust appearing as a construct in a relatively contained subset of studies. The supplementary rescan contributed 10 records (4.7% of the total screened pool of 215), confirming that keyword-based filtering alone would have missed a meaningful minority of trust-relevant records whose primary indexed terms did not foreground trust terminology.

4.2 OVERVIEW OF INCLUDED STUDIES

Table 1 presents the evidence matrix for the 10 included studies, summarizing key design characteristics and the normalized trust-target category for each study.

Table 1

Evidence matrix: included studies, design characteristics, and normalized trust-target category

Study	Authors (Year)	Journal	Design	N	Country	Trust Construct	Category
S1	Zhou; Lyu (2026)	Human Resources Management and Services	Scenario-based experiment	NR	China	AI trust as moderator of collaboration-to-innovation link	1
S2	Lin; Wang; Sheng (2025)	Humanities and Social Sciences Communications	Survey using TTF theory	302	China	Trust in GAI: willingness to rely on GAI	1
S3	Bales et al. (2025)	Frontiers in Neuroergonomics	Within-subject lab experiment	10	USA	Trust in robot agent (slider + EEG network metrics)	1
S4	Dawarka; Doargajudhur; Dutot (2026)	International Journal of Managing Projects in Business	Qualitative, interpretivist	NR	UK	Trust calibration in AI-enabled project teams	3
S5	Paparic; Bodea (2024)	Issues in Information Systems	Qualitative case study	NR	Ireland	Trust inside human project teams; GenAI as context	2
S6	Marmolejo-Ramos et al. (2024)	Frontiers in Artificial Intelligence	Indirect scenario experiment	3,260	20 countries	Trust in algorithmic decision-making (VAS)	1
S7	Naiseh et al. (2021)	World Wide Web	Qualitative (think-aloud, co-design)	16	NR	Trust calibration in AI recommendations	1
S8	Gay et al. (2019)	IEEE Access	Human-in-the-loop experiment (HLM)	32	USA	Operator suspicion toward AI-enabled system (SSI)	3
S9	Georganta; Ulfert (2024)*	Journal of Occupational and Organizational Psychology	Two pre-registered experiments	828	Germany / Netherlands	Interpersonal affective and cognitive trust in human-AI teams	2
S10	Niemann; Schmid (2025)	Electronic Communications of the EASST	Conceptual review	NR	NR	Trust, calibrated trust, and reliance in AI contexts	1

*Reclassified from raw Category 1 to Category 2 following full-text review; rationale detailed in Section 4.5. NR = Not reported. TTF = Technology-Task Fit. GAI = generative AI. VAS = Visual Analogue Scale. SSI = Situational Suspicion Inventory.
Source: Prepared by the author from full-text-confirmed studies indexed in Scopus.

The ten studies span publication years from 2019 to 2026, with a concentration in 2024 and 2025 (six studies). This temporal distribution reflects the acceleration of human-AI collaboration research following the mainstream adoption of large language model-based AI tools. In terms of design diversity, the corpus includes scenario-based experiments (S1, S6), a survey study (S2), a laboratory experiment (S3), qualitative and interpretivist studies (S4,

S5, S7), a human-in-the-loop experiment (S8), pre-registered between-subjects experiments (S9), and a conceptual review (S10). Sample sizes range from 10 (S3) to 3,260 (S6), with four studies not reporting a participant count (S1, S4, S5, S10). Geographic coverage spans China, the United States, the United Kingdom, Ireland, Germany, the Netherlands, and a 20-country international sample, though the depth of coverage across geographies varies substantially.

4.3 TRUST-TARGET DISTRIBUTION

Table 2 presents the frequency distribution of normalized trust-target categories.

Table 2

Frequency distribution of normalized trust-target categories (n = 10)

Category	Description	Studies (n)	Percentage
1	Trust in AI	6	60%
2	Interpersonal trust mediated by AI	2	20%
3	Ambiguous or unspecified	2	20%

Source: Prepared by the authors.

The distribution reveals a clear dominance of Category 1 operationalizations. Six of the ten studies treat the AI system as the primary trust object, operationalizing trust as the human user's willingness to rely on AI outputs or as the calibration of expectations about AI performance. Only two studies examine interpersonal trust among human team members as the primary dependent construct, with AI presence functioning as a structural variable. The remaining two studies present trust operationalizations that cannot be unambiguously assigned to either primary category. These findings directly address RQ1 and RQ2: the modal operationalization in human-AI collaboration research is Category 1 (trust in the AI system), and only a minority of studies distinguish interpersonal trust as a separate category.

4.4 CATEGORY 1 STUDIES: TRUST DIRECTED AT THE AI SYSTEM

S1. Zhou and Lyu (2026) examine the joint influence of employee-AI collaboration intensity and AI trust on employee innovative behavior in Chinese organizational contexts, using a scenario-based experiment. The independent variables are the level of employee-AI collaboration (high versus low) and the intensity of AI trust (high versus low), and the dependent variable is self-rated innovative behavior. The key finding is that employee-AI

collaboration positively affects innovative behavior through the mediating pathway of AI self-efficacy, and AI trust significantly moderates this mediated effect in a moderated mediation structure. Specifically, at high AI trust levels, the indirect effect of collaboration on innovation through self-efficacy is substantially stronger (effect coefficient = 0.2494) than at low levels (effect coefficient = 0.1017). AI trust is operationalized as “the degree to which employees are willing to rely on AI systems,” a canonical Category 1 definition. The study locates trust at the intersection of human factors (trust as a relational attitude toward the AI system) and organizational behavior (trust as a boundary condition for AI-enabled creative performance).

The study's contribution extends beyond the direct moderation finding. By treating AI trust as a moderator rather than a simple antecedent, Zhou and Lyu (2026) implicitly acknowledge that trust in AI is not the endpoint of the collaboration process but a governance parameter that shapes the benefit employees extract from their AI partnerships. This framing aligns with the calibration model of Lee and See (2004): the study distinguishes between high-trust and low-trust conditions to show that their practical consequences differ, without explicitly labeling this difference as calibration. A limitation noted by the authors is reliance on self-reported measures for both the mediating and dependent variables, raising the possibility of common method bias. The absence of a named psychometric trust scale in the extracted record is also noted (see Section 4.7).

S2. Lin, Wang and Sheng (2025) contribute a survey study (n = 302) examining the relationship between trust in generative AI and two forms of employee innovation: exploitative innovation (incremental refinement of existing processes) and exploratory innovation (development of genuinely novel approaches). Drawing on Technology-Task Fit (TTF) theory, the study proposes and finds that the relationship between trust in GAI and exploitative innovation is positive and linear, while the relationship with exploratory innovation is an inverted U shape (curvilinear): moderate trust is most conducive to exploratory behavior, with both very high and very low trust levels associated with reduced exploration. Employee-GAI fit and employee-innovation task fit are examined as moderators.

Trust in GAI is defined as “an individual's willingness to rely on GAI, based on the belief that it will perform tasks beneficially,” placing this squarely in Category 1. The curvilinear finding is theoretically rich and has practical implications for trust calibration: it suggests that maximizing trust in AI is not an unambiguous organizational objective, because at very high trust levels, the individual's critical engagement with AI outputs may be suppressed in ways that reduce exploration. This is consistent with the overtrust (misuse) pathway in Lee and

See's (2004) calibration model. The study limitation of single-country (China) sampling and self-reported measures is noted by the authors and represents a standing boundary on generalizability.

S3. Bales and colleagues (2025) investigate real-time trust inference in human-autonomy teaming through a within-subject laboratory experiment (n = 10) in which participants operate a UR5e robotic arm under two accuracy conditions (90% placement accuracy versus 30% placement accuracy). Trust is assessed through two simultaneous channels: a continuous behavioral self-report (a horizontal slider from 0 = no trust to 1 = complete trust, updated by the participant in real time during task execution) and EEG network metrics derived from inter-channel functional connectivity. The primary finding is that EEG network metrics outperform traditional EEG signal power metrics in predicting self-reported trust levels, with root mean squared error reductions ranging from 55% to 67% for some participants.

This study makes a methodological contribution to Category 1 trust research by demonstrating the technical feasibility of continuous trust inference from neurophysiological signals, which could support the development of adaptive automation systems that dynamically adjust their behavior based on real-time estimates of operator trust. The trust object is the robotic agent (Category 1), and the measurement instrument (the slider) is a behavioral self-report that captures real-time trust dynamics rather than a post-hoc survey-based summary. The study's limitation of a very small sample (n = 10) producing personalized within-subject models is explicitly acknowledged; the practical generalizability of the EEG-based approach will require validation with larger and more diverse samples.

S6. Marmolejo-Ramos and colleagues (2024) present the largest and most geographically diverse study in the corpus (n = 3,260 across 20 countries), examining trust in algorithmic decision-making through an indirect scenario-based experiment. Participants read a scenario in which an algorithm had made a decision (either a high-stakes or a low-stakes decision) and responded on a Visual Analogue Scale measuring their tendency to trust, recommend, or use the algorithm in that decision context. Explainability (with versus without an explanation of the algorithm's reasoning) was manipulated orthogonally to the stakes condition.

The key findings are threefold: statistical literacy has opposing effects depending on stakes (increasing trust in low-stakes AI decisions but decreasing it in high-stakes AI decisions); explainability alone does not significantly affect trust; and older participants and

male participants report lower trust in algorithms. These findings converge with and extend prior single-country research by establishing their replicability across a broad international sample. The trust object is the AI algorithm (Category 1), measured through the VAS scale. The study's design is cross-sectional and relies on self-reports via rating scales, acknowledged as limitations by the authors.

S7. Naiseh and colleagues (2021) address trust calibration in AI recommendation systems through a qualitative study in healthcare contexts (n = 16), using a think-aloud protocol to observe clinicians interacting with an AI decision support system, followed by co-design sessions to develop design principles for improved trust calibration. The trust target is AI recommendations (Category 1), assessed indirectly through the think-aloud protocol that captures whether and how participants engage with system explanations. The study's primary contribution is taxonomic and design-oriented: it identifies classes of trust calibration errors and derives five design principles for XAI systems. Because no numeric trust scale is administered, the “instrument” is the qualitative behavioral trace produced by the think-aloud protocol, documented as such in the reconciliation (see Section 4.7).

S10. Niemann and Schmid (2025) contribute a conceptual review of trust frameworks applicable to AI-assisted code generation, examining the theoretical and operational distinction between trust (an attitude) and reliance (a behavior). The paper maps competing definitions of trust from human factors, organizational psychology, and human-computer interaction, and argues that the apparent contradiction between high stated trust and low observed reliance (or vice versa) can be resolved by maintaining strict conceptual separation between the two constructs. The trust target in the frameworks reviewed is the AI coding assistant (Category 1). This is not an empirical study and collects no primary data, but its contribution as a conceptual clearing exercise is directly relevant to the measurement challenges identified in RQ3.

4.5 CATEGORY 2 STUDIES: INTERPERSONAL TRUST MEDIATED BY AI

S5. Paparic and Bodea (2024) conduct a qualitative case study in an Irish energy sector company examining how responsible usage of generative AI (specifically Microsoft Copilot) affects trust inside human project teams. The study identifies three main categories of responsible GenAI usage that have trust implications: strategic orientation toward responsible AI deployment, internal policy development that governs AI use in projects, and explicit trust-building actions among team members in the context of AI tool introduction. The

study also identifies inadequate AI usage, negative impacts on project workflow, and competence deficits as drivers of trust-related disruption within teams.

The key structural feature that assigns this study to Category 2 is the directionality of the trust construct: the dependent variable is trust between human team members within project teams (“trust inside the project teams”), and the GenAI tool is the contextual variable whose responsible or irresponsible usage shapes this interpersonal dynamic. The AI is not the trust object; the AI's presence and usage patterns are the conditions under which interpersonal trust is built or eroded. This distinguishes Paparic and Bodea (2024) from Category 1 studies in which trust is directed at the AI system itself.

The qualitative case study design does not yield a psychometric measurement of interpersonal trust; trust is assessed through thematic analysis of semi-structured interviews and organizational documents. While this limits numerical comparability with quantitative trust studies, it provides rich contextual detail about the organizational and social mechanisms through which AI presence affects interpersonal trust dynamics, mechanisms that would be invisible to psychometric measurement instruments. The study also flags an important practical limitation: some respondents had limited visibility into the extent of GenAI usage in the organization, which may have introduced blind spots in the organizational trust picture assembled from interview data.

S9. Georganta and Ulfert (2024) provide the only experimental study in the Category 2 subset, and their classification required a reclassification from the raw extraction value. In the extraction database, this study had been assigned a value associated with Category 1, reflecting a surface reading of the study as addressing AI-directed trust. The study's title (“Would you trust an AI team member?”) reinforces this initial impression. Full-text review, however, reveals a fundamentally different research architecture.

Both pre-registered experiments (Experiment 1: n = 385 dyadic teams in Germany; Experiment 2: n = 443 individuals in three-member teams across Germany and the Netherlands) use team composition as the independent variable, contrasting human-AI teams (in which one team member is a confederate simulating an AI teammate via the Wizard-of-Oz methodology) with human-human teams of equivalent size. The dependent variables in both experiments are interpersonal trust subscales: affective interpersonal trust (emotional bonding and confidence in the human teammate) and cognitive interpersonal trust (confidence in the human teammate's competence and reliability), measured from the

perspective of each human team member. Team identification and team trust are also assessed.

The findings show that, in two-member teams (Experiment 1), interpersonal trust mediated by perceived trustworthiness is significantly lower in human-AI teams than in human-human teams. The mechanism is perceived trustworthiness: AI team members are perceived as less trustworthy than human teammates, which translates into lower interpersonal trust reported by the human participants toward their non-AI teammates in the team. This spillover effect is the conceptual core of the Category 2 construct: the AI's membership in the team changes the interpersonal trust dynamics among the human members. In three-member teams (Experiment 2), the effect on overall team trust through the interpersonal trust pathway does not reach statistical significance, suggesting that team size may moderate the interpersonal spillover mechanism.

The reclassification of Georganta and Ulfert (2024) from Category 1 to Category 2 is supported by three converging pieces of evidence. First, the dependent variable is interpersonal trust between human team members, not trust directed at the AI team member. Second, the experimental manipulation is team composition (the presence or absence of an AI member), which is an independent variable that modifies the interpersonal context, not a target of the trust measurement. Third, the theoretical framework invoked by the authors explicitly draws on interpersonal trust and social identity theory, not on human-automation trust calibration models.

The reclassification has a material effect on the frequency distribution. Without the reclassification, Category 2 would contain only one study (10%), which could be attributed to idiosyncratic sampling or noise. With the reclassification, Category 2 contains two studies (20%), confirming that interpersonal trust mediated by AI is an independently identifiable construct present in the human-AI collaboration literature, even if substantially underrepresented relative to Category 1. This distinction is substantively important, as discussed in Section 4.8.

4.6 CATEGORY 3 STUDIES: AMBIGUOUS OR UNSPECIFIED TRUST TARGET

S4. Dawarka, Doargajudhur and Dutot (2026) present a qualitative study of trust calibration in AI-enabled project teams using an interpretivist research approach under a social constructivist epistemological framework. The study's central finding is that trust calibration in AI systems functions as a sociotechnical mechanism shaping both relational

and functional dimensions of team work, and that this mechanism influences collaboration quality and team performance in AI-enabled project environments. The study uses the term “trust calibration in AI systems” as its central construct.

The ambiguity that justifies Category 3 classification is structural. The interpretivist, constructivist design means that the trust construct emerges from the data rather than being operationalized in advance; the trust object is not specified by the researcher but is constituted through the participants' sensemaking. The resulting construct appears to encompass trust in AI technology, trust between team members in the context of AI use, and trust in the overall sociotechnical arrangement, without decomposing these into separate components with distinct measurement targets. Assigning Dawarka and colleagues (2026) to Category 1 or Category 2 would require imposing a directional operationalization that the study itself does not adopt. Category 3 therefore represents a faithful rather than a residual classification.

S8. Gay and colleagues (2019) study operator suspicion and team performance in a human-in-the-loop experiment in which military operators (n = 32) oversee an autonomous unmanned ground vehicle system under mission scenarios that include simulated cyber-attacks. The primary measurement instrument is the Situational Suspicion Inventory (SSI), which captures the operator's cognitive and behavioral orientation toward an automated system following an anomalous event. Suspicion, as operationalized through the SSI, refers to a heightened monitoring and verification disposition triggered by uncertainty about whether the system is operating as expected.

The assignment to Category 3 reflects the SSI's theoretical position relative to the trust construct. Suspicion is not equivalent to low trust or distrust; it is a qualitatively distinct state involving active investigative monitoring of system behavior that is not captured by conventional trust scales (which typically ask about willingness to rely, perceived competence, or confidence in system performance). The study finds that operator suspicion has a significant negative relationship with human-machine team performance (response score) and a positive relationship with response time, suggesting that heightened suspicion introduces a performance cost through increased verification behavior. This is a theoretically valuable finding about the trust-adjacent space, but the use of suspicion as the primary construct rather than a conventional trust measure means that the study does not permit direct assignment to Category 1 (which would require trust in the AI system as its dependent variable) or Category 2 (which would require interpersonal trust as its dependent variable).

The finding that “operator suspicion about AI system integrity negatively predicts team performance” is an important complement to Category 1 trust findings, suggesting that trust-adjacent states warrant dedicated treatment in future reviews.

4.7 INSTRUMENT RECONCILIATION

Table 3 presents the instrument reconciliation results across all 10 included studies.

Table 3

Instrument reconciliation: trust measurement across included studies

Study	Authors (Year)	Category	Condition	Instrument / Method	Notes
S1	Zhou; Lyu (2026)	1	II (Inferred)	Not named; scenario experiment design indicates probable validated scale	Extraction gap
S2	Lin; Wang; Sheng (2025)	1	II (Inferred)	Not named; large survey (n = 302) indicates probable validated scale	Extraction gap
S3	Bales et al. (2025)	1	I (Documented)	Horizontal slider (0 to 1) under On Board Camera panel; EEG network metrics for inference	Behavioral continuous measure + neurophysiological inference
S4	Dawarka et al. (2026)	3	III (N/A)	Qualitative interpretivist; no psychometric instrument applicable	By design
S5	Paparic; Bodea (2024)	2	III (N/A)	Qualitative case study; no psychometric instrument applicable	By design
S6	Marmolejo-Ramos et al. (2024)	1	I (Documented)	VAS rating scale (0 = not at all likely, 5 = very likely)	Behavioral VAS measure
S7	Naiseh et al. (2021)	1	I (Documented)	Think-aloud protocol as behavioral trust trace	Qualitative behavioral trace
S8	Gay et al. (2019)	3	III (N/A)	Situational Suspicion Inventory (SSI)	Trust-adjacent; not a trust scale
S9	Georganta; Ulfert (2024)	2	II (Inferred)	Not named; pre-registered experiments indicate probable validated interpersonal trust scales	Extraction gap
S10	Niemann; Schmid (2025)	1	III (N/A)	Conceptual review; no primary data collected	By design

Condition I = Instrument documented; Condition II = Instrument inferred (extraction gap); Condition III = No psychometric instrument applicable.
Source: Prepared by the author.

The reconciliation reveals three distinct instrument patterns.

Pattern A (Documented, non-standard): Three studies (S3, S6, S7) document their trust measurement procedures explicitly, but all three employ non-traditional formats rather

than named multi-item psychometric scales. The Bales and colleagues (2025) study uses a continuous slider updated in real time, which constitutes a behavioral measure rather than a reflective self-report instrument. The Marmolejo-Ramos and colleagues (2024) study uses a VAS scale embedded in the experimental scenarios, again capturing a behavioral tendency rather than administering a validated trust inventory. The Naiseh and colleagues (2021) study uses a think-aloud protocol to generate a qualitative behavioral trace. None of the three provides a multi-item validated scale by name (such as the Trust in Automation scale, the Organizational Trust Inventory, or a domain-specific validated instrument), which would allow cross-study comparison of trust levels at a standardized psychometric level.

Pattern B (Inferred from design; extraction gap): Three studies (S1, S2, S9) present designs strongly consistent with the use of validated psychometric instruments but do not disclose the instrument name in the extracted record. Zhou and Lyu (2026) use a scenario-based experiment with AI trust as a moderating variable at two levels; the experimental treatment of trust at quantitative thresholds (effect coefficients cited to four decimal places) implies that trust was measured through a validated continuous scale, not merely dichotomized. Lin, Wang and Sheng (2025) administer a survey to 302 participants using a theorized model in which trust in GAI is a key variable; survey-based empirical work of this design type routinely employs published validated scales. Georganta and Ulfert (2024) conducted two pre-registered experiments, and the pre-registration standard in organizational and occupational psychology typically requires specifying the measurement instruments in advance. The absence of instrument names in these three abstracted records is therefore classified as an extraction artifact rather than a genuine signal of instrument absence.

Pattern C (No psychometric instrument applicable): Four studies (S4, S5, S8, S10) genuinely lack a psychometric trust measurement instrument by design. Dawarka and colleagues (2026) and Paparic and Bodea (2024) use qualitative methods for which psychometric instruments are methodologically inappropriate. Gay and colleagues (2019) use the SSI, an instrument designed to capture a trust-adjacent construct (suspicion) rather than trust itself. Niemann and Schmid (2025) conduct a conceptual review with no primary data collection.

The aggregate pattern addresses RQ3: among studies that do deploy psychometric or behavioral trust instruments, the specific instruments are often not captured in the abstracted record, constituting an infrastructure gap in systematic review practice. Future studies and

review protocols should treat instrument documentation (scale name, item count, response format, and reliability coefficient) as a mandatory rather than optional extraction field.

4.8 THE NEAR-ABSENCE OF CATEGORY 2 AS A SUBSTANTIVE FINDING

The finding that only two of the ten included studies operationalize interpersonal trust mediated by AI presence (Category 2) deserves direct interpretive attention. This near-absence is not a methodological artifact of the review process. The eligibility criteria required only that a trust construct be present in the context of human-AI collaboration, without specifying its direction; the supplementary rescan extended the search beyond the primary prefilter; and the reclassification of Georganta and Ulfert (2024) from the extraction database's Category 1 coding to Category 2 adds one additional study that would otherwise have been miscounted as Category 1. Given this inclusive architecture, the relative scarcity of Category 2 studies is a genuine empirical signal about the current structure of the literature.

Several factors help explain this pattern. First, the intellectual lineage of trust research in human-computer interaction and human-automation research has historically framed trust as a property of the dyadic human-technology relationship, directing both theoretical attention and measurement instrumentation toward the technology as the trust object (LEE; SEE, 2004; HANCOCK *et al.*, 2011). This framing has become embedded in research design norms: researchers approaching human-AI collaboration from a human factors or human-computer interaction perspective are more likely to ask “how much does the user trust the AI system?” than “how does the AI's presence as a team member change the trust dynamics among the human teammates?” The former question can be addressed with well-established research protocols; the latter requires a more complex design.

Second, the study of interpersonal trust in organizational settings has developed largely as a separate tradition within organizational and occupational psychology (MAYER; DAVIS; SCHOORMAN, 1995). As AI agents have begun to enter team compositions, researchers from this tradition have started to ask Category 2 questions, as Georganta and Ulfert (2024) illustrate. However, this research stream is in an early phase of formation, and its integration with the human factors literature on Category 1 trust remains methodologically underdeveloped.

Third, Category 2 research designs present substantially greater methodological complexity than Category 1 designs. To study interpersonal trust mediated by AI, a researcher must (a) manipulate AI presence as an independent variable affecting intra-team dynamics,

(b) measure interpersonal trust as the dependent variable using an instrument designed for human-human trust rather than human-machine trust, (c) control for or separately assess Category 1 trust to prevent confound, and (d) ensure that the AI-presence manipulation does not simply alter participants' beliefs about the task rather than the social composition of their team. The Wizard-of-Oz design used by Georganta and Ulfert (2024) addresses challenge (a) by simulating an AI team member with a human confederate, but this design introduces its own validity concerns about whether participants' responses reflect trust in AI team members or trust in unfamiliar human collaborators whose behavior has been made to appear algorithmically constrained.

The practical significance of the near-absence of Category 2 research is considerable. As AI agents are integrated into collaborative work in organizations at increasing scale, the effects of AI team membership on human interpersonal trust within those teams constitute an undertheorized and underresearched domain. If AI presence systematically reduces interpersonal trust (as the Georganta and Ulfert, 2024, dyadic findings suggest for two-member teams), then organizations deploying AI teammates may be inadvertently degrading the social fabric of their human teams while monitoring only the adoption metrics associated with Category 1 trust.

Moreover, the Category 1 and Category 2 trust constructs are not independent. As Dellermann and colleagues (2019) imply in the hybrid intelligence framework, a human team member who trusts the AI's recommendations highly may delegate more decision authority to the AI, reducing coordination with human teammates and potentially eroding the interpersonal trust structures that effective team coordination requires. The mechanisms linking Category 1 and Category 2 trust dynamics remain largely unexplored in the literature, and their investigation will require designs that simultaneously measure both trust targets.

4.9 RESEARCH GAPS

The analysis of the ten included studies identifies four priority research gaps for the human-AI collaboration and trust literature.

Gap 1 (Category 2 underrepresentation): The most critical gap is the empirical underrepresentation of Category 2 research. Future work should explicitly design studies in which interpersonal trust between human team members is the primary dependent variable and AI presence (or AI-presence intensity, or AI-to-human ratio in the team) is the independent variable, using validated interpersonal trust instruments and team settings with

adequate experimental control. The Georganta and Ulfert (2024) study provides a methodological template, including the pre-registration standard and the Wizard-of-Oz implementation approach for teams with simulated AI members.

Gap 2 (Instrument standardization and reporting): The instrument-reporting gap identified in Section 4.7 signals a systematic deficit in measurement documentation in the human-AI trust literature. Future studies should report the full instrument specification (scale name, item count, response format, Cronbach's alpha or other reliability estimate, and the publication from which the scale was drawn) as standard documentation. Journals and systematic review protocols should treat this information as mandatory rather than optional, enabling future reviewers to assess measurement equivalence across studies.

Gap 3 (Cultural and sectoral diversity): Of the ten included studies, eight are from China, the United States, the United Kingdom, Ireland, or a cross-national sample dominated by Western contexts. Marmolejo-Ramos and colleagues (2024) provide valuable cross-national data but focus exclusively on Category 1 trust. The Category 1 versus Category 2 distinction may manifest differently across high-collectivism and high-individualism cultural contexts: in high-collectivism cultures, where team cohesion and interpersonal relational quality may be more salient to team members, the disruption of interpersonal trust by AI team membership (Category 2) may have larger practical consequences than in high-individualism contexts, where individual reliance on AI tools (Category 1) is more likely to be the primary trust dynamic.

Gap 4 (Longitudinal dynamics): All ten included studies employ cross-sectional or experimental point-in-time designs. Trust in human-AI collaboration is theoretically expected to evolve over extended interaction periods: initial trust before AI experience is likely to be based on reputation and general disposition to trust technology; trust during early AI use is likely to be highly sensitive to initial performance experiences; and trust in established long-term human-AI collaborations may be more stable and resistant to recalibration. Longitudinal designs tracking trust trajectories, and specifically the relative dynamics of Category 1 and Category 2 trust over time as teams gain experience working with AI members, are entirely absent from the corpus.

5 CONCLUSION

This systematic review set out to map how human-AI collaboration research operationalizes the trust-target construct, to assess the degree to which the literature

distinguishes trust directed at AI systems from interpersonal trust modulated by AI presence, and to characterize the measurement instruments used for each trust target. Applying a PRISMA 2020 search and screening protocol to a Scopus-sourced corpus, and conducting a supplementary rescan to ensure recall coverage, the review identified and synthesized ten full-text-confirmed studies that met the eligibility criteria.

The central finding is that the dominant operationalization of trust in human-AI collaboration research treats the AI system as the primary trust object (Category 1 = 60%). Trust directed at human team members in contexts where AI presence is the modifying factor (Category 2 = 20%) remains substantially underrepresented. Two studies present trust operationalizations that cannot be unambiguously classified and are assigned to Category 3 (20%). These proportions directly answer RQ1 and RQ2: the field has a strong and relatively uniform tendency to frame trust as a human-to-AI relational attitude, and the interpersonal dimension of trust in AI-inclusive teams has yet to develop as an independently represented research area.

A secondary finding concerns instrument reporting (RQ3). Of the ten included studies, only three document a clearly named measurement procedure; four use non-psychometric or qualitative approaches appropriate to their designs; and three show extraction gaps where validated instruments were likely used but not captured in the abstracted record. The last pattern is particularly concerning from a cumulative science standpoint: if instruments are not documented, subsequent reviewers cannot assess whether nominally convergent findings across studies reflect convergent validity in the trust construct or are artifacts of instrument-specific variance. The finding that named psychometric instruments are not systematically documented in this literature is an independent methodological finding of the review, distinct from the trust-target analysis.

The near-absence of Category 2 studies is interpreted as a substantive empirical finding about the current state of the literature rather than as a shortcoming of the review design. It reflects a structural feature of the field's theoretical grammar: the dominant frameworks for studying trust in automation and human-computer interaction were developed before AI agents began to participate in human teams as social actors, and those frameworks direct conceptual and measurement attention toward the human-system dyad rather than toward the interpersonal consequences of AI team membership. As AI systems transition from peripheral tools to embedded team members, the social and relational architecture of teams

that include AI agents requires theoretical frameworks and research designs capable of addressing both trust targets simultaneously.

5.1 THEORETICAL CONTRIBUTIONS

The three-category coding scheme introduced in this review makes a conceptual precision contribution to the study of trust in sociotechnical systems. It operationalizes at the level of research coding a distinction that theoretical frameworks have implied but not formalized: the distinction between trust in an AI system (a human-to-technology attitude) and interpersonal trust modulated by AI presence (a human-to-human attitude shaped by the social and structural context that the AI creates). For researchers working within the Mayer, Davis and Schoorman (1995) framework, the scheme clarifies that the ability-benevolence-integrity triad applies to different targets in Category 1 and Category 2 studies: in Category 1, the triad describes the AI system; in Category 2, it describes the human team member whose trustworthiness is being assessed in the context of an AI-modified team composition.

The scheme also extends the O'Neill and colleagues (2022) call for conceptual clarity in human-autonomy teaming research by providing an operational coding instrument applicable to any corpus of trust studies, making the clarification cumulable across reviews and replicable by independent coders.

5.2 PRACTICAL CONTRIBUTIONS

Organizations deploying AI systems in collaborative work settings typically assess AI adoption through surveys of employee trust in the AI technology. The present findings indicate that such assessments capture only the Category 1 dimension of trust dynamics in AI-augmented teams. The effects of AI team membership on the interpersonal trust among human team members (Category 2) are not captured by these instruments, meaning that organizations may be deploying AI teammates without visibility into the second-order social consequences of those deployments for team cohesion and collaboration quality.

The instrument-reconciliation findings have implications for review infrastructure: systematic review protocols for human-AI collaboration research should incorporate instrument documentation as a mandatory extraction field to enable measurement-level comparisons across studies.

5.3 LIMITATIONS

This review has four principal limitations. First, the search is restricted to the Scopus database. Human-AI collaboration research is distributed across publication channels including conference proceedings, preprint repositories, and domain-specific databases that Scopus may index selectively. Second, the corpus of ten studies is sufficient for the categorical analysis conducted here but is too small to support statistical meta-analysis or to estimate effect sizes. Third, the supplementary rescan increases recall but cannot guarantee exhaustiveness; studies in which trust is entirely implicit or is addressed under terminologically distinct labels may remain outside the corpus. Fourth, the trust-target coding scheme requires interpretive judgments at the boundaries between categories, particularly for ambiguous cases assigned to Category 3.

5.4 FUTURE RESEARCH DIRECTIONS

Priority directions for future work include: (i) the development of experimental designs that position AI presence as the independent variable with interpersonal trust as the primary dependent variable, using validated interpersonal trust instruments in controlled team settings; (ii) the standardization of instrument documentation requirements in journals publishing human-AI trust research; (iii) cross-cultural comparative studies examining whether the Category 1 versus Category 2 balance differs across collectivistic and individualistic national contexts; and (iv) longitudinal studies tracking the co-evolution of Category 1 and Category 2 trust as human teams gain sustained experience working alongside AI team members.

The three-category framework advanced in this review is offered as a living instrument: as the human-AI collaboration literature matures and new forms of AI team integration emerge, the coding criteria may require refinement, and additional categories may become necessary to accommodate trust constructs that current designs have not yet produced. The framework's value lies not in its finality but in its capacity to make the trust-target dimension of human-AI collaboration research visible, comparable, and cumulable across studies.

ACKNOWLEDGMENTS

The author thanks the institutional library services of Universidade Estadual de Montes Claros (Unimontes) for database access supporting the systematic search.

Funding: The sources of funding for this research, if any, require confirmation by the author prior to submission. Any institutional, governmental, or private funding received in support of this work must be disclosed, including grant numbers and funder names, in accordance with the target journal's funding transparency policy.

Competing interests: The presence or absence of financial and non-financial competing interests requires confirmation by the author. If no competing interests exist, this should be affirmatively declared.

Data availability: The systematic review source data, including screening decision records, extracted fields, trust-target normalization coding sheets, and instrument reconciliation records, require disclosure of their availability and access conditions. If the data are to be deposited in an open repository, the repository name, accession number, and access conditions should be stated here.

Code availability: If computational scripts were used in the search, screening, or normalization process, their availability and access conditions require confirmation and disclosure.

Pre-registration: The pre-registration status of this systematic review protocol requires confirmation. If the protocol was registered prior to data collection (for example, via PROSPERO or the Open Science Framework), the registration number and date should be cited here. If the review was not pre-registered, this should be affirmatively stated.

REFERENCES

BALES, G.; HAYMAN, A. P. A.; CLARK, T. K.; DEKARSKE, J.; JOSHI, S.; KONG, Z. An EEG-network-metric based approach to real-time trust inference in human-autonomy teaming. Frontiers in Neuroergonomics, v. 6, art. 1627483, 2025. DOI: 10.3389/fnrgo.2025.1627483.

DAWARKA, V.; DOARGAJUDHUR, M.; DUTOT, V. AI in project teams: how trust calibration reconfigures team's collaboration and performance. International Journal of Managing Projects in Business, p. 1-22, 2026. DOI: 10.1108/IJMPB-07-2025-0285.

DELLERMANN, D.; EBEL, P.; SÖLLNER, M.; LEIMEISTER, J. M. Hybrid intelligence. Business and Information Systems Engineering, v. 61, n. 5, p. 637-643, 2019. DOI: 10.1007/s12599-019-00595-2.

GAY, C.; HOROWITZ, B.; ELSHAW, J. J.; BOBKO, P.; KIM, I. Operator suspicion and human-machine team performance under mission scenarios of unmanned ground vehicle operation. IEEE Access, v. 7, p. 36371-36379, 2019. DOI: 10.1109/ACCESS.2019.2901258.

GEORGANTA, E.; ULFERT, A.-S. Would you trust an AI team member? Team trust in human-AI teams. Journal of Occupational and Organizational Psychology, v. 97, n. 3, p. 1212-1241, 2024. DOI: 10.1111/joop.12504.

HANCOCK, P. A.; BILLINGS, D. R.; SCHAEFER, K. E.; CHEN, J. Y. C.; DE VISSER, E. J.; PARASURAMAN, R. A meta-analysis of factors affecting trust in human-robot interaction. Human Factors, v. 53, n. 5, p. 517-527, 2011. DOI: 10.1177/0018720811417254.

LEE, J. D.; SEE, K. A. Trust in automation: designing for appropriate reliance. Human Factors, v. 46, n. 1, p. 50-80, 2004. DOI: 10.1518/hfes.46.1.50_30392.

LIN, X.; WANG, T.; SHENG, F. Exploring the dual effect of trust in GAI on employees' exploitative and exploratory innovation. Humanities and Social Sciences Communications, v. 12, n. 1, art. 663, 2025. DOI: 10.1057/s41599-025-04956-z.

MARMOLEJO-RAMOS, F. *et al.* Factors influencing trust in algorithmic decision-making: an indirect scenario-based experiment. Frontiers in Artificial Intelligence, v. 7, art. 1465605, 2024. DOI: 10.3389/frai.2024.1465605.

MAYER, R. C.; DAVIS, J. H.; SCHOORMAN, F. D. An integrative model of organizational trust. Academy of Management Review, v. 20, n. 3, p. 709-734, 1995. DOI: 10.5465/amr.1995.9508080335.

McNEESE, N. J.; DEMIR, M.; COOKE, N. J.; MYERS, C. Teaming with a synthetic teammate: insights into human-autonomy teaming. Human Factors, v. 60, n. 2, p. 262-273, 2018. DOI: 10.1177/0018720817743223.

NAISEH, M.; AL-THANI, D.; JIANG, N.; ALI, R. Explainable recommendation: when design meets trust calibration. World Wide Web, v. 24, n. 5, p. 1857-1884, 2021. DOI: 10.1007/s11280-021-00916-0.

NIEMANN, S.; SCHMID, U. Human-AI collaboration in coding: a trust perspective. Electronic Communications of the EASST, v. 84, 2025. DOI: 10.14279/eceasst.v84.2679.

O'NEILL, T. A.; McNEESE, N.; BARRON, A.; SCHELBLE, B. Human-autonomy teaming: a review and analysis of the empirical literature. Human Factors, v. 64, n. 5, p. 904-938, 2022. DOI: 10.1177/0018720820960865.

PAGE, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, v. 372, n. 71, 2021. DOI: 10.1136/bmj.n71.

PAPARIC, M.; BODEA, C.-N. Building trust through responsible usage of generative artificial intelligence in projects: a case study. Issues in Information Systems, v. 25, n. 4, 2024. DOI: 10.48009/4_iis_2024_112.

Received: [date to be confirmed upon submission] | Accepted: [date to be confirmed upon submission] | Published: [date to be confirmed upon submission]

SEEBER, I. *et al.* Machines as teammates: a research agenda on AI in team collaboration. Information and Management, v. 57, n. 2, art. 103174, 2020. DOI: 10.1016/j.im.2019.103174.

ZHOU, Y.; LYU, B. Employee-AI collaboration and employee innovative behavior: the roles of AI self-efficacy and AI trust. Human Resources Management and Services, v. 8, n. 2, 2026. DOI: 10.18282/hrms5880.