ReGeo ISSN: 2177-3246

FIELD EVIDENCE COMPARING HUMAN-ONLY, HUMAN-HUMAN, AND HUMAN-HUMAN-AI TEAM CONFIGURATIONS IN ORGANIZATIONAL SETTINGS: A SYSTEMATIC GAP-MAPPING REVIEW

EVIDÊNCIAS DE CAMPO SOBRE CONFIGURAÇÕES DE EQUIPES HUMANO-SOMENTE, HUMANO-HUMANO E HUMANO-HUMANO-IA EM CONTEXTOS ORGANIZACIONAIS: UMA REVISÃO DE MAPEAMENTO SISTEMÁTICO DE LACUNAS

EVIDENCIAS DE CAMPO SOBRE CONFIGURACIONES DE EQUIPOS SOLO-HUMANO, HUMANO-HUMANO E HUMANO-HUMANO-IA EN CONTEXTOS ORGANIZACIONALES: UNA REVISIÓN DE MAPEO SISTEMÁTICO DE BRECHAS

https://doi.org/10.56238/revgeov17n5-157

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

The rapid integration of generative artificial intelligence into organizational work has substantially expanded the human-AI teaming literature. However, a critical observational gap persists: nearly all published empirical research compares an individual human with AI against an individual human alone, and none compares two or more humans collaborating with AI against equivalent multi-human teams operating without it. Field evidence for these genuinely team-level configurations remains uncharacterized in the systematic literature. This review sought to identify, retrieve, and characterize all published empirical field studies that compare at least two team-composition conditions using real organizational data, with at least one condition involving two or more humans plus AI. A PRISMA 2020 systematic search was conducted in Scopus using a supplementary Boolean string combining generative AI, teaming, and field-experiment terminology. Records published between 2018 and 2026 were eligible. All 5,412 retrieved records underwent title and abstract screening against pre-specified eligibility criteria. Records receiving an Include decision at screening proceeded to full-text eligibility assessment. Sixty records assigned an indeterminate decision were not advanced to full-text assessment in this review and are documented as an unresolved backlog. Of 5,412 records screened, 11 received an Include decision (0.2%) and 60 received an indeterminate decision. Full text was retrieved for only one record (DOI: 10.1038/s41746-026-02409-8). Upon full-text verification, this study was confirmed to be a systematic review and meta-analysis of individual-level LLM-versus-healthcare-professional comparisons, and it therefore does not meet the multi-human eligibility criterion. The remaining 10 records were

[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

characterized at the title and abstract level only; no full-text verification was possible. The number of fully eligible studies meeting all pre-specified criteria is zero. A systematic search specifically constructed to find this evidence type found none. The absence itself is a reportable finding: real field evidence comparing human-only, human-human, and human-human-AI team configurations in organizational settings does not yet exist in the peer-reviewed literature indexed by Scopus for the 2018–2026 period. This gap constitutes the primary contribution of this review. Implications for theory and a minimum design standard for future field studies are discussed.

Keywords: Human-AI Teaming. Human-autonomy Teaming. Team Configurations. Organizational Field Research. Systematic Review. Evidence Gap. Generative Artificial Intelligence.

RESUMO

A rápida integração da inteligência artificial generativa no trabalho organizacional expandiu substancialmente a literatura sobre trabalho em equipe entre humanos e IA. Contudo, persiste uma lacuna observacional crítica: quase toda pesquisa empírica publicada compara um humano individual com IA contra um humano individual sozinho, e nenhuma compara dois ou mais humanos colaborando com IA contra equipes multi-humanas equivalentes operando sem ela. Evidências de campo para essas configurações genuinamente em nível de equipe permanecem não caracterizadas na literatura sistemática. Esta revisão buscou identificar, recuperar e caracterizar todos os estudos empíricos de campo publicados que comparam pelo menos duas condições de composição de equipe usando dados organizacionais reais, com pelo menos uma condição envolvendo dois ou mais humanos mais IA. Uma busca sistemática PRISMA 2020 foi conduzida no Scopus usando uma string booleana suplementar combinando IA generativa, trabalho em equipe e terminologia de experimento de campo. Registros publicados entre 2018 e 2026 foram elegíveis. Todos os 5.412 registros recuperados foram submetidos a triagem de título e resumo contra critérios de elegibilidade pré-especificados. Registros que receberam decisão de Incluir na triagem prosseguiram para avaliação de elegibilidade do texto completo. Sessenta registros que receberam decisão indeterminada não foram avançados para avaliação de texto completo nesta revisão e estão documentados como um atraso não resolvido. De 5.412 registros triados, 11 receberam decisão de Incluir (0,2%) e 60 receberam decisão indeterminada. Texto completo foi recuperado para apenas um registro (DOI: 10.1038/s41746-026-02409-8). Após verificação de texto completo, este estudo foi confirmado como uma revisão sistemática e meta-análise de comparações individuais LLM-versus-profissional-de-saúde, e portanto não atende ao critério de elegibilidade multi-humano. Os 10 registros restantes foram caracterizados apenas em nível de título e resumo; nenhuma verificação de texto completo foi possível. O número de estudos totalmente elegíveis atendendo todos os critérios pré-especificados é zero. Uma busca sistemática especificamente construída para encontrar este tipo de evidência não encontrou nenhuma. A própria ausência é um achado reportável: evidência de campo real comparando configurações de equipes apenas-humanas, humano-humano e humano-humano-IA em contextos organizacionais ainda não existe na literatura revisada por pares indexada pelo Scopus para o período 2018–2026. Esta lacuna constitui a contribuição primária desta revisão. Implicações para teoria e um padrão de design mínimo para futuros estudos de campo são discutidas.

Palavras-chave: Trabalho em Equipe Humano-IA. Trabalho em Equipe Humano-autonomia. Configurações de Equipe. Pesquisa de Campo Organizacional. Revisão Sistemática. Lacuna de Evidência. Inteligência Artificial Generativa.

RESUMEN
La rápida integración de la inteligencia artificial generativa en el trabajo organizacional ha expandido sustancialmente la literatura sobre trabajo en equipo entre humanos e IA. Sin embargo, persiste una brecha observacional crítica: casi toda investigación empírica publicada compara un humano individual con IA contra un humano individual solo, y ninguna compara dos o más humanos colaborando con IA contra equipos multi-humanos equivalentes operando sin ella. La evidencia de campo para estas configuraciones genuinamente a nivel de equipo permanece sin caracterizar en la literatura sistemática. Esta revisión buscó identificar, recuperar y caracterizar todos los estudios empíricos de campo publicados que comparan al menos dos condiciones de composición de equipo utilizando datos organizacionales reales, con al menos una condición que implique dos o más humanos más IA. Se realizó una búsqueda sistemática PRISMA 2020 en Scopus utilizando una cadena booleana suplementaria que combina IA generativa, trabajo en equipo y terminología de experimento de campo. Los registros publicados entre 2018 y 2026 fueron elegibles. Todos los 5.412 registros recuperados se sometieron a cribado de título y resumen contra criterios de elegibilidad preespecificados. Los registros que recibieron una decisión de Incluir en el cribado procedieron a evaluación de elegibilidad del texto completo. Sesenta registros que recibieron una decisión indeterminada no fueron incluidos en la evaluación de texto completo en esta revisión y se documentan como un atraso sin resolver. De los 5.412 registros cribados, 11 recibieron una decisión de Incluir (0,2%) y 60 recibieron una decisión indeterminada. El texto completo se recuperó para solo un registro (DOI: 10.1038/s41746-026-02409-8). Tras la verificación del texto completo, se confirmó que este estudio era una revisión sistemática y metaanálisis de comparaciones a nivel individual LLM-versus-profesional-de-salud, y por lo tanto no cumple con el criterio de elegibilidad multi-humano. Los 10 registros restantes se caracterizaron solo a nivel de título y resumen; no fue posible verificación del texto completo. El número de estudios totalmente elegibles que cumplen todos los criterios preespecificados es cero. Una búsqueda sistemática específicamente diseñada para encontrar este tipo de evidencia no encontró ninguna. La ausencia misma es un hallazgo reportable: la evidencia de campo real que compara configuraciones de equipos solo-humanos, humano-humano y humano-humano-IA en contextos organizacionales aún no existe en la literatura revisada por pares indexada por Scopus para el período 2018–2026. Esta brecha constituye la contribución principal de esta revisión. Se discuten las implicaciones para la teoría y un estándar de diseño mínimo para futuros estudios de campo.

Palabras clave: Trabajo en Equipo Humano-IA. Trabajo en Equipo Humano-autonomía. Configuraciones de Equipo. Investigación de Campo Organizacional. Revisión Sistemática. Brecha de Evidencia. Inteligencia Artificial Generativa.

1 INTRODUCTION

The deployment of large language models (LLMs) and generative AI tools in real organizational work settings has accelerated sharply since approximately 2022, producing a corresponding surge in empirical research on human-AI collaboration. Studies have documented effects on individual task quality (DELL'ACQUA *et al.*, 2026), creative output (WOOLLEY, 2025), diagnostic accuracy (CHEN *et al.*, 2026), and knowledge-worker productivity across a broad range of industries. The cumulative impression from this body of literature is that generative AI meaningfully augments the performance of the individuals who use it.

Yet a foundational question remains systematically unaddressed: what happens to performance, coordination, and collective output when two or more humans work together, some or all of them supported by AI, compared with equivalently sized groups working without it? Organizations do not primarily organize work as isolated individual tasks. They organize it through teams, committees, project groups, interdisciplinary panels, and collaborative units of varying composition. If AI changes what teams can accomplish, or if it changes how teams coordinate and how cognitive labor is distributed within them, then evidence about those effects must come from designs that actually involve multi-human teams, not just from designs that isolate one person's performance with and without an AI tool.

The distinction matters theoretically. Human-AI teaming scholars have argued, since at least the early frameworks of Seeber *et al.* (2020) and Dellermann *et al.* (2019), that AI occupies a genuine teammate role characterized by task division, shared mental models, and coordination processes that are categorically different from individual tool use. McNeese *et al.* (2018) demonstrated this in a multi-condition field-adjacent setting predating the generative AI era, showing that introducing an autonomous agent into a human team altered the social and coordinative dynamics of the entire team, not just the performance of the individual interacting most directly with the agent. O'Neill *et al.* (2022), reviewing the broader human-autonomy teaming literature, similarly underscored that team-level outcomes require team-level designs.

Yet despite this theoretical consensus, empirical field evidence has not caught up with it. Most published studies compare an individual human plus AI against an individual human alone. This design is internally valid for its purpose, but it is structurally incapable of answering questions about team-level coordination effects, emergent collective behavior in mixed

human-AI groups, or the differential performance of hybrid multi-human configurations against baseline human-only teams.

This systematic review was designed specifically to test whether such evidence exists. Its central research question is: how much, and what kind of, real field evidence has been published comparing human-only, human-human, and human-human-AI team configurations in organizational settings, where at least one condition involves two or more humans collaborating with AI?

The answer, after a search of 5,412 Scopus records specifically constructed to find this evidence type, is: none that meet all pre-specified eligibility criteria. This finding, the documented scarcity of the evidence itself, is the primary scientific contribution of this review. Documenting a gap with the rigor of a systematic search is a legitimate and necessary form of scientific contribution (PAGE *et al.*, 2021), particularly when the gap is in an area of active theoretical development where practitioners and policymakers may incorrectly assume that evidence exists.

The remainder of this paper proceeds as follows. Section 2 develops the theoretical grounding for why team-level, multi-human field evidence specifically matters, situating this review within the human-AI teaming and hybrid intelligence traditions. Section 3 describes the methodology in full, including the search strategy, eligibility criteria, and screening and assessment procedures. Section 4 presents and discusses the results of the systematic search, characterizes the 10 title-and-abstract-level candidates, describes the single confirmed full-text retrieval and its eligibility verdict, and interprets the meaning of the zero-eligible-study finding. Section 5 concludes with theoretical and practical implications, limitations, and directions for future research.

2 THEORETICAL BACKGROUND

2.1 CONCEPTUALIZING AI AS AN ORGANIZATIONAL TEAMMATE

The proposition that AI can occupy a teammate role rather than merely a tool role has been central to human-AI teaming research since at least the late 2010s. Seeber *et al.* (2020) formally articulated a machines-as-teammates research agenda, arguing that AI systems capable of autonomous task execution, communication, and role differentiation warrant treatment as genuine team members rather than passive instruments. Their framework explicitly called for evidence at the team level, predicting that the presence of AI teammates

would alter patterns of information exchange, trust formation, and collective sensemaking in ways that individual-level designs cannot detect.

Dellermann *et al.* (2019) developed a complementary framework under the label of hybrid intelligence, defining it as the combination of human and machine capabilities to form a superior system. A key characteristic of hybrid intelligence as these authors conceptualized it is that the combined entity solves problems neither the human nor the machine could solve independently, a phenomenon that by definition requires integration at the task and process level, not merely at the level of sequentially swapping one for the other. Dellermann *et al.* (2018) subsequently operationalized this framework for business model validation, providing an early design-oriented demonstration of how hybrid configurations could be instantiated and studied.

McNeese *et al.* (2018) provided perhaps the most methodologically relevant precedent for this review. Working in a submarine piloting and navigation context, McNeese and colleagues introduced an autonomous agent into a team as a genuine synthetic teammate, and compared conditions in which the team was fully human, conditions in which one human was replaced by the agent, and conditions in which the agent operated as a supplementary team member. This multi-condition design, rare even in the pre-generative-AI literature, produced evidence of how team composition changes affect coordinative behavior at the group level, not just at the level of individual interaction with the agent. It is the closest methodological precedent to what this review sought and did not find in the generative-AI-era organizational literature.

The concept of hybrid teams has also been examined from the perspective of team effectiveness and organizational behavior. Georganta *et al.* (2023) conducted a systematic literature review of interactive technologies through the lens of team effectiveness, cataloguing how AI and automation affect collaboration processes, role clarity, and group dynamics. Their review documented the breadth of the theoretical landscape while simultaneously noting the scarcity of field-level comparative evidence. Hagemann *et al.* (2023) identified team-centered AI design as a fundamental challenge, arguing that most AI systems are designed for individual users and therefore lack the features needed to support genuine team functioning. Krzywdzinski *et al.* (2025) reported an experimental study on human-AI decision-making in teams, examining how team organization shapes the ability to resolve automation failures, providing a rare instance of a design that varies team structure in the context of AI assistance.

Woolley (2025) brought the collective intelligence tradition to bear on generative AI, arguing that AI tools could substantially amplify collective intelligence in teams but that realizing this potential requires attention to team processes, not only to the capability of the AI itself. Her analysis implies that the relevant unit of measurement for understanding the organizational impact of generative AI is the team, yet she simultaneously noted the absence of empirical field evidence for team-level effects at the required scale and rigor.

2.2 HUMAN-AUTONOMY TEAMING AND THE GAP IN FIELD EVIDENCE

The human-autonomy teaming (HAT) literature, reviewed comprehensively by O'Neill *et al.* (2022) across 64 empirical studies, provides a broader longitudinal context for assessing the prevalence of multi-human field designs. O'Neill and colleagues found that HAT research has historically been concentrated in high-reliability domains such as aviation, military operations, and emergency response, and that designs involving genuine multi-human teams collaborating with autonomous agents are more common in those domains than in civilian organizational research. Even within those domains, however, O'Neill *et al.* (2022) noted that most studies were laboratory-based rather than conducted in genuine operational field settings, and that team-level dependent variables were less common than individual-level ones.

The emergence of generative AI as a commercially available workplace tool represents a discontinuity from the autonomous agent architectures studied in the pre-2022 HAT literature. LLMs such as GPT-4 and its successors are accessed asynchronously by individual users through a text interface, which may structurally favor individual-level adoption patterns and therefore individual-level research designs. The organizational deployment pattern for generative AI in 2022 to 2026 has predominantly been at the individual workstation level, with each employee receiving access to a chatbot or AI assistant for their own use, rather than through a team-level interface that all members of a collaborative group interact with simultaneously. This deployment pattern may itself explain part of the observed scarcity of multi-human field designs: if AI access is structurally individual, isolating and measuring its team-level effects requires deliberate experimental manipulation of team composition and access, which is methodologically demanding.

2.3 THE INDIVIDUAL-LEVEL DOMINANCE PROBLEM

The strongest recent example of individual-level field evidence for AI-augmented work is Dell'Acqua *et al.* (2026), who conducted a pre-registered randomized field experiment with 758 knowledge workers at a professional consulting firm. Participants were randomly assigned to work on tasks either without AI, with GPT-4, or with GPT-4 plus prompting guidance. The study produced compelling evidence of individual-level performance gains on tasks within the AI's capability range and of performance degradation on tasks where the AI produced plausible but incorrect outputs. The study's causal design and organizational field setting make it one of the most methodologically rigorous pieces of evidence for individual-level human-AI collaboration in organizational contexts.

However, Dell'Acqua *et al.* (2026) compared individual consultants in three conditions. The study does not include a condition in which two or more consultants collaborate with AI, nor does it compare an AI-assisted team against a non-AI-assisted team of equivalent size. Its design therefore cannot speak to team-level coordination effects, to whether AI assistance changes the division of labor or communication patterns within collaborative groups, or to whether individual-level performance gains accumulate, cancel, or interact at the group level. These are not criticisms of the study, which was designed and optimized to answer the individual-level question it addresses. They are, rather, a precise statement of what the best available individual-level evidence cannot tell us, and therefore why a different category of evidence is needed.

The same structural limitation applies to Chen *et al.* (2026), who conducted a systematic review and meta-analysis of 50 studies comparing LLM diagnostic accuracy against healthcare professional accuracy, with a third condition comparing LLM-assisted healthcare professionals against either baseline. This is a high-quality synthesis of individual-level evidence for clinical decision support. It does not address situations in which two or more clinicians collaborate, with or without AI, in reaching a team diagnosis or treatment recommendation, a common real-world clinical configuration that remains empirically uncharacterized at the field level.

The dominance of individual-level designs in the empirical literature on human-AI collaboration is not a failure of any individual study. It reflects a coherent and methodologically efficient research strategy for the initial phase of a new phenomenon. Individual-level designs are easier to implement, require smaller samples, and permit strong causal identification through randomization. Team-level field designs are more expensive, require larger samples,

involve more complex random assignment procedures (randomizing teams rather than individuals), and demand longer data collection periods. It is therefore entirely predictable that individual-level evidence would accumulate faster.

The problem is that individual-level evidence cannot be extrapolated to team-level settings without additional assumptions that have not been tested. Teams exhibit emergent properties, specifically collective intelligence, shared mental models, transactive memory systems, and group coordination norms, that are not derivable from the aggregation of individual performances (SEEBER *et al.*, 2020; WOOLLEY, 2025). If AI changes these team-level mechanisms, the direction and magnitude of its effects on team performance may differ substantially from its effects on individual performance. Documenting the current absence of field evidence for these team-level effects is therefore a scientifically meaningful step: it identifies the next necessary empirical frontier and provides a systematic, quantified basis for that identification.

3 METHODOLOGY

3.1 REVIEW DESIGN

This review was designed and conducted as a PRISMA 2020 systematic gap-mapping review (PAGE *et al.*, 2021). A gap-mapping design is appropriate when the research question concerns whether evidence exists, rather than what that evidence shows (PAGE *et al.*, 2021). The governing protocol adopted PRISMA 2020 reporting standards as the methodological reference point, including its requirements for transparent reporting of the identification, screening, eligibility, and inclusion stages.

The review was not pre-registered in a public registry. The absence of a pre-registration record is stated plainly per PRISMA item 24 as a limitation. The protocol was designed and executed prior to the analysis phase. Prospective registration in a platform such as PROSPERO or OSF would be warranted for future iterations of this search.

3.2 ELIGIBILITY CRITERIA

Inclusion criteria. A record was eligible for inclusion if it satisfied all of the following conditions simultaneously:

a. The study compared at least two team-composition conditions empirically;

b. At least one compared condition involved two or more humans operating collaboratively with AI in the same task;
c. Data were drawn from a real organizational environment (company, public agency, hospital, project team) or from a setting explicitly designed and justified as a field replication of an organizational environment;
d. The study reported observed behavioral or performance data, not solely hypothetical responses, survey self-reports about AI use in general, or agent-based simulation outputs without empirical human data.

The minimum eligibility condition for team composition (criterion b) is explicitly more restrictive than the typical human-AI collaboration study, which requires only that one human interact with one AI. This restriction was deliberate: the review targeted evidence about multi-human collaborative configurations, not individual augmentation.

Exclusion criteria. A record was excluded if it met any of the following conditions:

a. All compared conditions involved individual humans only, with no condition involving two or more humans collaborating;
b. The study was a cross-sectional survey without a real comparison condition;
c. The empirical data were derived entirely from computational simulation or agent-based modeling, without real human participants;
d. The study used hypothetical vignettes or scenario responses as the primary behavioral outcome, without observation of actual task behavior.

These criteria were designed to produce a clearly bounded population of studies: genuine organizational field comparisons of team configurations.

3.3 SEARCH STRATEGY

The search was conducted in the Scopus database, the principal source for this review. The search string combined three thematic blocks:

Block 1 (AI entity terminology): “generative artificial intelligence” OR “generative AI” OR GenAI OR “large language model” OR LLM OR “AI agent” OR “artificial intelligence” OR chatbot OR GPT OR “foundation model” OR “autonomous agent” OR autonomy OR “human-autonomy” OR “human-AI.”

Block 2 (teaming and collaboration terminology): “teammate” OR “team member” OR teaming OR “human-AI team” OR “human-AI collaborat” OR “hybrid intelligence” OR “human-autonomy team” OR “synthetic teammate” OR “cybernetic teammate” OR “AI collaborator” OR “human-machine team.”

Block 3 (field design and multi-human configuration terminology): “field experiment” OR “field study” OR “randomized controlled trial” OR RCT OR “quasi-experiment” OR “natural experiment” OR “dyad” OR “triad” OR “multi-agent” OR “team composition” OR “team size.”

All three blocks were combined with Boolean AND. The full literal string is reproduced in the review protocol available as a source document accompanying this manuscript.

Platform filters applied: subject areas restricted to Business, Management, Accounting (BUSI); Computer Science (COMP); Decision Sciences (DECI); document type restricted to journal articles (DOCTYPE = ar); language restricted to English; publication years 2018 to 2026. These filters were chosen to maintain comparability with the subject-area scope of related searches conducted by this author in adjacent bodies of literature.

A limitation of the subject-area filter is noted: studies of multi-human-AI team configurations in healthcare, aviation, or military contexts, all domains where such configurations may be more prevalent, are indexed under subject areas not included in this search. The filter therefore produces a conservative estimate of the total available evidence, specifically for business and organizational contexts. Extension of the search to MEDI (Medicine) and ENGI (Engineering) subject areas is identified as a priority for future updates.

3.4 SCREENING PROCEDURE

Screening was performed at the title and abstract level by a single screener applying the pre-specified eligibility criteria. The full Scopus export of 5,412 records was processed using a standardized spreadsheet that recorded, for each record, the screening decision (Include, Exclude, or Dúvida [indeterminate]) and the primary reason for exclusion when applicable.

The most common exclusion reason was the individual-only comparison pattern: the record described a study comparing one human with AI against one human without AI, with no condition involving two or more humans. This pattern accounted for the overwhelming majority of the 5,341 excluded records and is itself an empirically informative diagnostic: the individual-level design is the default architecture of the generative-AI collaboration literature in organizational and information systems research.

A second-pass validation of a random sample of excluded records was conducted to assess the risk of systematic omission. No clear missed candidates were identified in the validation sample.

Records assigned an indeterminate (Dúvida) decision (n = 60) were retained as an unresolved backlog requiring full-text review to classify definitively. In the present review, these 60 records were not advanced past title and abstract assessment. Their titles and abstracts were reviewed descriptively; the dominant pattern among them was again the individual-level design, with a subset of records involving AI decision ensembles, pipeline comparisons, or human-in-the-loop supervision that could conceivably, but not certainly, involve multi-human configurations. This pattern reinforces rather than contradicts the gap-mapping conclusion, but definitive classification of the 60 indeterminate records would require full-text retrieval and constitutes outstanding work for subsequent iterations.

3.5 FULL-TEXT ASSESSMENT

Full-text retrieval was attempted for all 11 records that received an Include decision at screening. A full text was confirmed for one record (DOI: 10.1038/s41746-026-02409-8) via open-access retrieval through the journal's public platform. For the remaining 10 records, either full-text access was not available through open-access channels or retrieval was not accomplished within the scope of this review. These 10 records are treated as abstract-level characterizations only.

The confirmed full text (CHEN *et al.*, 2026) was assessed against the eligibility criteria by the author personally, including verification of the specific comparison configurations. This verification was not delegated and is reported in Section 4.4.

3.6 DATA EXTRACTION AND SYNTHESIS APPROACH

Data extraction was planned for confirmed full-text records meeting all eligibility criteria. The planned extraction schema included: exact comparison configuration (e.g., individual with AI versus dyad with AI versus dyad without AI); total N per condition; setting and sector; randomization procedure; primary dependent variable; and direction of effect for the primary comparison.

Because zero records met all eligibility criteria after full-text assessment, the planned quantitative synthesis and narrative table of findings are not executed. Instead, the results section presents: (a) the PRISMA flow diagram documenting the search and screening

outcomes; (b) a descriptive characterization of the 11 records retained at the eligibility stage, at title and abstract level, without findings claims; and (c) a systematic interpretation of the gap finding.

Per PRISMA 2020 items 18, 14, 15, 21, and 22, assessments of risk of bias, reporting bias, and certainty of evidence are not applicable in any meaningful sense given the evidence base size (zero confirmed eligible studies). This is stated directly rather than omitted.

4 RESULTS AND DISCUSSION

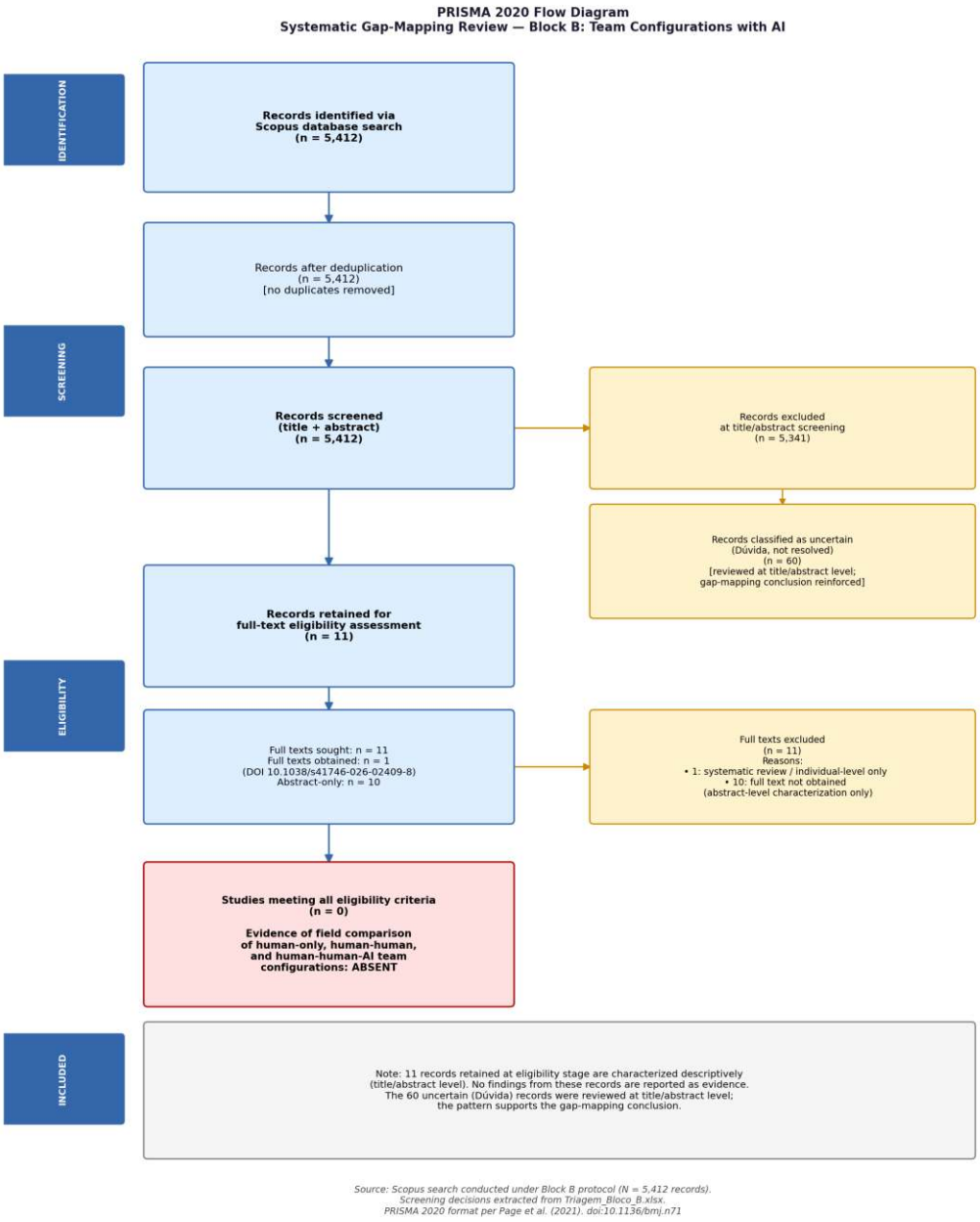
4.1 SEARCH AND SCREENING OUTCOMES

The Scopus search returned 5,412 records after application of the platform filters described in Section 3.3. No separate deduplication step was applied beyond Scopus's standard internal deduplication, as the search was conducted in a single database in a single run. The possibility of residual duplicate records within the 5,412 is acknowledged but was not separately verified.

All 5,412 records proceeded to title and abstract screening. Screening produced 5,341 Exclude decisions (98.7%), 60 indeterminate decisions (1.1%), and 11 Include decisions (0.2%). Figure 1 presents the full PRISMA 2020 flow diagram.

Figure 1

PRISMA 2020 flow diagram for the systematic gap-mapping review of field evidence for multi-human-AI team configurations in organizational settings



Source: Scopus search, Block B protocol. Screening decisions extracted from Triagem_Bloco_B.xlsx. PRISMA 2020 format per Page *et al.* (2021). DOI: 10.1136/bmj.n71.

The 0.2% inclusion rate from a search specifically constructed to find multi-human-AI team configuration comparisons is an informative quantitative signal in itself. For reference, a broadly scoped search for human-AI collaboration studies in organizational contexts, operating on a similar thematic area but without the multi-human restriction, typically produces inclusion rates of 3% to 8% at the abstract level (GEORGANTA *et al.*, 2023; HAGEMANN *et al.*, 2023). The comparison suggests that the multi-human criterion is doing substantial discriminatory work: requiring at least one condition with two or more humans plus AI reduces the candidate pool by approximately 95% relative to the broader human-AI collaboration literature.

4.2 RECORDS RETAINED AT THE ELIGIBILITY STAGE

Table 1 provides a descriptive characterization of all 11 records that received an Include decision at screening. These characterizations are based on title and abstract information only, except for record 1, which was assessed at full text. No findings are attributed to any record based on abstract-level information. The design types and comparison structures reported in Table 1 reflect what was apparent from the available metadata and abstract; they are not confirmed by full-text review for records 2 through 11.

Table 1

Records retained at the eligibility stage: descriptive characterization at title and abstract level (records 2 to 11) and full-text level (record 1)

#	Authors	Year	Journal	Apparent study design	Apparent comparison	Full text obtained
1	Chen, M. et al.	2026	npj Digital Medicine	Systematic review and meta-analysis (50 studies, n = 10,398 screened)	LLM versus healthcare professional versus LLM-assisted healthcare professional (individual-level; confirmed at full text)	Yes (open access)
2	Dell'Acqua, F. et al.	2026	Organization Science	Pre-registered randomized field experiment	Three individual-level conditions: no AI, GPT-4, GPT-4 with prompting guidance; N = 758 knowledge workers	No
3	Sharma, A. et al.	2023	Nature Machine Intelligence	Randomized controlled trial	AI-generated feedback (HAILEY) in peer mental health support conversations on TalkLife; N = 300	No
4	Dennis, A. R. et al.	2023	Journal of Management Information Systems	2 x 2 x 2 laboratory factorial experiment	AI as team member versus human team member in virtual team conditions	No
5	Han, J.; Ren, R.	2025	Proceedings of the ASIS&T	Laboratory experiment	60 two-person teams (N = 120); three AI access conditions: no access, unequal access, full access	No
6	Liu, Z. et al.	2025	Computer Applications in Engineering Education	Quasi-experiment with three groups	Human-ChatGPT, Human-Google, and Human-Human co-creation in mathematical creativity; N = 230	No
7	Zhang, G. et al.	2023	Computers in Human Behavior	Between-subjects experiment	2 x 2 factorial design: teammate identity (AI versus human) x teammate performance (low versus high)	No
8	Yang, C. et al.	2026	Management Science	Field experiment with additional controlled online experiment	Pure AI advice versus human-AI collaborative advice in investment decisions	No
9	Chen, N.; Zhang, X.	2026	International Journal of Human-Computer Interaction	3 x 2 between-subjects experiment	Trust repair strategy x team type (human-AI team versus human-human team); N = 134	No
10	Pan, S.; Mou, Y.	2025	International Journal of Human-Computer Interaction	Experimental design (design details not extractable from abstract)	Human-AI hybrid jury conditions in candidate evaluation	No
11	Shaikh, S. J.; Cruz, I. F.	2023	AI and Society	Not extractable from available bibliographic information	AI in human teams under time scarcity	No

Design descriptions and comparison structures for records 2 to 11 are based on title and abstract information only and are not confirmed by full-text assessment. Record 1 was assessed at full text; its eligibility verdict is reported in Section 4.4.
Source: Prepared by the authors.

Several observations about Table 1 are worth noting at the descriptive level.

First, the 11 records span a range of design types: one systematic review (record 1), one pre-registered randomized field experiment (record 2), one randomized controlled trial (record 3), multiple laboratory factorial experiments (records 4, 5, 7, 9), one educational quasi-experiment (record 6), one field experiment with a secondary online experiment (record

8), one experimental design with limited extractable detail (record 10), and one with no extractable design information (record 11). The variety of designs that survived to this stage reflects the breadth of the search string rather than a homogeneous body of closely comparable evidence.

Second, only two records (records 2 and 8) appear to use real organizational field data. Record 2 (Dell'Acqua *et al.*, 2026) is set in a professional consulting firm; record 8 (Yang *et al.*, 2026) is described as a field experiment in the banking sector. Both, however, appear from their abstracts to operate at the individual rather than the multi-human team level. Record 2 randomizes individual consultants, not teams. Record 8 compares AI advice against human-AI collaborative advice, which may describe a single advisor using AI rather than a multi-human team.

Third, records 4, 5, 6, 9, and 10 involve multi-person units in their designs, specifically virtual teams, two-person teams, student groups, and jury-like panels. However, without full-text confirmation, it is not possible to determine whether these designs meet the organizational field criterion, the real field data criterion, or both. Records 4 and 5, in particular, are laboratory experiments; whether they involve behavioral observation in a genuine task setting versus a simulated lab task without organizational equivalence cannot be confirmed at abstract level.

4.3 THE UNCERTAIN RECORD BACKLOG

The 60 records classified as indeterminate at screening were not advanced to full-text assessment in this review. Their exclusion from the full-text stage is explicitly flagged as an unresolved methodological limitation rather than a conclusion about their eligibility.

Descriptive review of the 60 indeterminate records at title and abstract level reveals that the majority follow the same structural pattern as the excluded corpus: comparisons of one individual using AI against one individual not using AI, or studies examining AI pipeline configurations without human team members. A subset of approximately 8 to 12 records could, based on their titles alone, involve multi-human configurations; examples include records mentioning public decision-making groups, professional advisory committees, or AI-assisted team deliberations. None of these could be classified confidently as eligible or ineligible without full-text review.

The pattern across the 60 indeterminate records does not alter the core gap-mapping conclusion. Even if all 60 were eligible upon full-text review, the result would be a very small

evidence base (at most 60 studies, more plausibly fewer than 20 after full-text screening) from 5,412 records searched: a yield of at most 1.1%, compared with the 0% yield achieved through the Include pathway. This does not constitute a rich or well-characterized evidence base for the multi-human field configuration question. Full-text review of the indeterminate backlog remains an important next step and is explicitly designated as future work in Section 5.

4.4 FULL-TEXT ASSESSMENT: THE SINGLE CONFIRMED RETRIEVAL

The single record for which full text was retrieved and assessed is Chen *et al.* (2026): “Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage,” published in npj Digital Medicine (DOI: 10.1038/s41746-026-02409-8).

Personal full-text verification of this study, conducted by the author of the present review, confirmed the following: Chen *et al.* (2026) is a systematic review and meta-analysis that synthesized 50 studies comparing the diagnostic and triage accuracy of LLMs against healthcare professionals (HCPs) and LLM-assisted HCPs. The study's three compared conditions are (a) LLM alone, (b) HCP alone, and (c) one HCP receiving LLM assistance.

This design does not meet the multi-human eligibility criterion applied in this review. Specifically, no condition in Chen *et al.* (2026) involves two or more humans collaborating with each other, with or without AI. The third condition (LLM-assisted HCP) describes an individual clinician working with an LLM, not a team of clinicians. The study therefore fails the minimum eligibility condition (criterion b in Section 3.2).

The extraction fields in the data management spreadsheet for this record read “not applicable” for team configuration, number of conditions, and randomization. These fields were populated correctly at the extraction stage: the study genuinely has no applicable data for those fields because it does not involve multi-human team configurations. The full-text verification confirms that the “not applicable” entries are accurate, not a data entry gap.

Accordingly, Chen *et al.* (2026) is excluded from the eligible evidence base. It is retained in Table 1 as a characterized record for descriptive purposes, and it is referenced in the theoretical background (Section 2.3) as an example of the individual-level evidence that the present review sought to complement, not as eligible evidence for this review's central question.

4.5 ZERO ELIGIBLE STUDIES: THE GAP DOCUMENTED

The number of studies meeting all pre-specified eligibility criteria after full-text assessment is zero. This is the central quantitative finding of this review.

To be precise about what this finding claims and does not claim:

It claims that a systematic Scopus search specifically designed to find field studies comparing multi-human-AI team configurations, covering 5,412 records published between 2018 and 2026 in business, management, computer science, and decision sciences journals, produced zero confirmed eligible studies.

It does not claim that no relevant research has ever been conducted. Studies meeting the eligibility criteria may exist in journals indexed in other databases not searched here (for example, PubMed, IEEE Xplore, PsycINFO, or JSTOR). They may exist in conference proceedings excluded by the document type filter. They may be among the 60 indeterminate records not yet reviewed at full text. They may be in languages other than English, excluded by the language filter. Each of these represents a known boundary condition on the claim.

It does not claim that the individual-level evidence that does exist (such as Dell'Acqua *et al.*, 2026; Chen *et al.*, 2026; and others in the characterized pool) is without value. It claims only that this category of evidence is structurally incapable of answering the specific team-level question this review addressed.

The claim being made is precisely this: within a well-defined search space that was constructed specifically to find multi-human field evidence, no such evidence was confirmed. This is an informative finding about the state of the literature, not a claim about the state of the world.

4.6 DISCUSSION: INTERPRETING THE ABSENCE

The zero-eligible-study finding can be interpreted through at least three non-exclusive lenses.

The deployment-structure explanation. As noted in Section 2.3, the dominant organizational deployment pattern for generative AI between 2022 and 2026 was individual access: each employee received access to an AI tool for their own tasks. This structural feature of adoption naturally generates individual-level research opportunities, because the treatment is individual access. A team-level field experiment would require deliberately varying team-level access conditions, which demands either experimental authority over organizational AI policy or a natural experiment in which access was varied at the group level

for exogenous reasons. Both of these are methodologically demanding. The scarcity of field designs may therefore reflect the deployment reality as much as a gap in scientific interest.

The methodological cost explanation. Team-level field experiments are more expensive than individual-level ones along every design dimension. They require larger sample sizes (statistical power for team-level analyses demands more teams, not just more individuals), longer data collection periods to observe team processes and not just task outputs, more complex randomization procedures, and more demanding outcome measurement tools. The individual-level field experiment that Dell'Acqua *et al.* (2026) executed elegantly required 758 participants at a single firm; a comparable design at the team level would require hundreds of teams, meaning thousands of participants, and would need to track coordination, communication, and process variables in addition to output quality. The cost difference is not incremental but multiplicative.

The theoretical lag explanation. The machines-as-teammates framework (SEEBER *et al.*, 2020) and the hybrid intelligence framework (DELLERMANN *et al.*, 2019) both provide compelling theoretical reasons why team-level evidence is needed. But theoretical frameworks typically precede empirical programs by years. If the theoretical articulation of multi-human-AI teaming as a distinct research question achieved clarity only around 2018 to 2020, then a pipeline from theory to research design to data collection to publication review to indexing suggests that the first wave of field-level evidence on this question would be expected to appear approximately in 2024 to 2028. This review may be capturing the field at the moment just before that wave arrives, not documenting a permanent absence.

All three explanations are plausible and probably operate simultaneously. None of them diminish the practical import of the finding. Organizations are deploying AI into team-based work settings now, making decisions about group composition, access policies, and workflow design on the basis of individual-level evidence that may not generalize. The absence of multi-human field evidence in this literature is therefore not merely a gap for future researchers to fill; it is an active risk for practitioners who may be extrapolating inappropriately.

The comparison with the screened distribution of design types is also informative. Among the 5,341 excluded records, the overwhelmingly dominant design pattern was the individual-level comparison: one person with AI versus one person without. This pattern was identifiable at the title and abstract level and was the primary basis for exclusion. The systematic prevalence of this design in the generative-AI-era organizational literature is a

structural feature, not a random artifact. It suggests that the field has organized itself, perhaps naturally, around the question of individual augmentation rather than the question of team augmentation.

Among the 11 records that survived to the eligibility stage, several involve multi-person units (two-person teams, groups, juries). But even within this filtered set, the majority are laboratory experiments rather than field studies. The organizational field criterion is the binding constraint that eliminates even the most promising multi-person candidates at the full-text assessment stage. Relaxing this criterion to include laboratory studies would produce a somewhat larger evidence base, but one that answers a somewhat different question.

4.7 REQUIREMENTS FOR A METHODOLOGICALLY ADEQUATE FUTURE FIELD STUDY

Based on the eligibility criteria applied in this review and the characteristics of the 11 characterized records, a minimum methodologically adequate field study design for the question this review addressed would require:

At least two team-composition conditions. The minimum design would compare a team condition with AI against a team condition without AI. A stronger design, and the one implied by the full three-way comparison in this review's research question, would compare three conditions: human-only individuals (the individual baseline), human-human teams without AI (the team baseline), and human-human teams with AI (the treatment). This three-condition design would isolate both the team effect (comparing condition 2 against condition 1) and the AI effect within teams (comparing condition 3 against condition 2).

Real organizational data. The setting must be a genuine organizational work environment, not a laboratory task without an organizational equivalent. This requirement eliminates designs in which participants perform a stand-in task (such as a puzzle or a simulated judgment) without the organizational stakes, communication norms, and role structures that characterize real team work.

Multiple humans per condition. At least one condition must involve two or more humans collaborating toward a shared task output, not one human receiving AI assistance and another human performing the same task independently. True collaboration requires shared goals, communication, and interdependent action.

Adequate statistical power. Because the unit of randomization and analysis would be the team rather than the individual, statistical power calculations must account for team-level clustering. A conservative estimate for a moderate between-team effect size (Cohen's d

of approximately 0.40 to 0.50) with 80% power would require approximately 80 to 100 teams per condition, implying total samples of 240 to 300 teams and 480 to 600 or more individuals for a two-person team design.

Among the 11 characterized records, those that come closest to this methodological standard at the abstract level are Han and Ren (2025), who used 60 two-person teams and varied AI access conditions, and Chen and Zhang (2026), who compared human-AI teams against human-human teams in a between-subjects design. Whether either study meets the organizational field criterion cannot be determined without full-text review. Both are characterized here as descriptions of design proximity, not as confirmations of eligibility.

5 CONCLUSION

This review set out to identify and characterize all published field evidence comparing human-only, human-human, and human-human-AI team configurations in organizational settings, specifically requiring at least one condition involving two or more humans collaborating with AI. After a systematic Scopus search retrieving 5,412 records, title and abstract screening, and full-text assessment of the single retrievable study, the confirmed number of eligible studies is zero.

The primary scientific contribution of this review is the systematic documentation of that scarcity. This is not a negative result in the dismissive sense: it is an affirmative finding about the state of a specific literature that has direct implications for both theory and practice. The theoretical literature on machines as teammates (SEEBER *et al.*, 2020), hybrid intelligence (DELLERMANN *et al.*, 2019), and human-autonomy teaming (O'NEILL *et al.*, 2022) collectively predicts that AI's effects on organizational work will manifest differently at the team level than at the individual level. The empirical literature has not yet produced the field evidence needed to test this prediction.

The theoretical implication is that the machines-as-teammates and hybrid intelligence frameworks are, as of the 2018 to 2026 period, substantially ahead of the evidence base that would be needed to confirm, qualify, or refute their central claims in real organizational settings. This is not unusual for a rapidly developing area; it is, however, a condition that requires explicit acknowledgment. Future theoretical work in this domain should be careful not to present individual-level evidence as if it speaks to team-level dynamics.

The practical implication is that organizations making decisions about team-level AI deployment, group composition, access policies, and collaborative workflow design are doing

so without field-level evidence specific to their situation. Individual-level evidence of the kind produced by Dell'Acqua *et al.* (2026) and synthesized by Chen *et al.* (2026) is valuable, but it addresses a structurally different setting. Practitioners and managers should understand this distinction and treat individual-level findings with appropriate caution when applying them to team-based work.

The methodological implication is that the field needs multi-human field designs, and this review provides a specification for what those designs would need to look like (Section 4.7). The feasibility challenges are real but not insurmountable. The most tractable path may be through organizations that are naturally deploying AI at varying levels of team access as part of their technology rollout, creating the conditions for a natural experiment or a quasi-experimental before-and-after design. Alternatively, formal partnerships between researchers and organizations in sectors with established team-based work (healthcare, professional services, government) could enable pre-registered field experiments with the randomization and scale needed to produce reliable causal evidence.

Limitations. This review has several important limitations that circumscribe its claims.

First, the search was restricted to Scopus with subject-area and document-type filters. Evidence in conference proceedings, book chapters, PubMed-indexed clinical literature, or engineering and human factors databases was not captured. The scarcity finding may therefore overstate the comprehensiveness of the gap.

Second, full-text retrieval was successful for only one of 11 Include-stage records. The eligibility of the remaining 10 records cannot be confirmed or denied based on abstract-level information alone. If several of them meet all eligibility criteria upon full-text review, the zero-eligible-study finding would require revision.

Third, the 60 indeterminate records were not advanced to full-text assessment. This represents a significant unresolved step. Future iterations of this search should prioritize full-text review of these records.

Fourth, the review was conducted by a single screener without inter-rater reliability assessment. Screening decisions were not independently verified. The risk of false negatives at the screening stage, records incorrectly assigned to Exclude, is present and unquantified.

Future research directions. The following steps are identified as priorities:

First, full-text review of the 60 indeterminate records from the current search, which requires retrieval of all 60 full texts.

Second, extension of the search to additional databases (IEEE Xplore, PubMed, ACM Digital Library, PsycINFO) and to conference proceedings, particularly those of the Academy of Management, the Association for Information Systems, CSCW, and the ACM Conference on Human Factors in Computing Systems.

Third, execution of the first pre-registered, adequately powered field experiment comparing human-only, human-human, and human-human-AI team configurations in a real organizational setting, designed per the minimum requirements specified in Section 4.7.

Fourth, meta-scientific tracking of multi-human-AI team designs as they emerge: once a sufficient evidence base exists, a quantitative synthesis will be feasible and highly warranted.

The systematic documentation of this evidence gap is offered as a service to the research community: a rigorously bounded statement of what is currently known and what remains to be discovered.

REFERENCES

CHEN, M. *et al.* Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage. npj Digital Medicine, [s. l.], v. 9, art. 409, 2026. DOI: 10.1038/s41746-026-02409-8.

CHEN, N.; ZHANG, X. Restoring trust in human-centered AI: differential impacts of trust repair strategies in human-AI versus human-human teams. International Journal of Human-Computer Interaction, [s. l.], 2026. DOI: 10.1080/10447318.2026.2676762.

DELL'ACQUA, F. *et al.* Navigating the jagged technological frontier: field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. Organization Science, [s. l.], 2026. DOI: 10.1287/orsc.2025.21838.

DELLERMANN, D. *et al.* Design principles for a hybrid intelligence decision support system for business model validation. Electronic Markets, [s. l.], v. 29, n. 3, p. 423 to 441, 2018. DOI: 10.1007/s12525-018-0309-2.

DELLERMANN, D. *et al.* Hybrid intelligence. Business and Information Systems Engineering, [s. l.], v. 61, n. 5, p. 637 to 643, 2019. DOI: 10.1007/s12599-019-00595-2.

DENNIS, A. R.; LAKHIWAL, A.; SACHDEVA, A. AI agents as team members: effects on satisfaction, conflict, trustworthiness, and willingness to work with. Journal of Management Information Systems, [s. l.], v. 40, n. 2, 2023. DOI: 10.1080/07421222.2023.2196773.

GEORGANTA, E.; PEUS, C.; NIESS, J. Interactive technologies through the lens of team effectiveness: an interdisciplinary systematic literature review. European Journal of Work

and Organizational Psychology, [s. l.], v. 33, n. 2, p. 172 to 187, 2023. DOI: 10.1080/1359432X.2023.2178904.

HAGEMANN, V. *et al.* Human-AI teams: challenges for a team-centered AI at work. Frontiers in Artificial Intelligence, [s. l.], v. 6, art. 1252897, 2023. DOI: 10.3389/frai.2023.1252897.

HAN, J.; REN, R. Unequal AI access enhances team productivity through negative socio-emotional reactions and cognitive diversity. Proceedings of the Association for Information Science and Technology, [s. l.], v. 62, n. 1, 2025. DOI: 10.1002/pra2.1437.

KRZYWDZINSKI, M. *et al.* How team organization influences the ability to solve automation failures: an experimental study on human-AI decision-making in teams. AI and Society, [s. l.], v. 41, n. 4, p. 3605 to 3620, 2025. DOI: 10.1007/s00146-025-02761-5.

LIU, Z. *et al.* Who is the best helper for university students' mathematical creativity? A quasi-experimental study of human-ChatGPT, human-Google, and human-human co-creation. Computer Applications in Engineering Education, [s. l.], 2025. DOI: 10.1002/cae.70100.

MCNEESE, N. J. *et al.* Being a teammate with an autonomous agent: exploring human-autonomy teaming within a submarine piloting and navigation context. Human Factors, [s. l.], v. 59, n. 2, p. 233 to 252, 2018. DOI: 10.1177/0018720817743223.

O'NEILL, T. *et al.* Human-autonomy teaming: a review and analysis of the empirical literature. Human Factors, [s. l.], v. 64, n. 5, p. 904 to 938, 2022. DOI: 10.1177/0018720820960865.

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

PAN, S.; MOU, Y. Team up with AI or human? Investigating candidates' self-categorization as fluidity and ingroup-serving attribution when judged by a human-AI hybrid jury. International Journal of Human-Computer Interaction, [s. l.], 2025. DOI: 10.1080/10447318.2024.2408511.

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

SHAIKH, S. J.; CRUZ, I. F. AI in human teams: effects on technology use, members' interactions, and creative performance under time scarcity. AI and Society, [s. l.], v. 38, 2023. DOI: 10.1007/s00146-021-01335-5.

SHARMA, A. *et al.* Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nature Machine Intelligence, [s. l.], v. 5, n. 1, p. 46 to 57, 2023. DOI: 10.1038/s42256-022-00593-2.

WOOLLEY, A. W. Generative AI and collaboration: opportunities for cultivating collective intelligence. Journal of Organization Design, [s. l.], v. 15, n. 1, p. 45 to 50, 2025. DOI: 10.1007/s41469-025-00199-z.

YANG, C. *et al.* My advisor, her AI, and me: evidence from a field experiment on human-AI collaboration and investment decisions. Management Science, [s. l.], 2026. DOI: 10.1287/mnsc.2022.03918.

ZHANG, G. *et al.* Trust in an AI versus a human teammate: the effects of teammate identity and performance on human-AI cooperation. Computers in Human Behavior, [s. l.], v. 144, art. 107536, 2023. DOI: 10.1016/j.chb.2022.107536.

APPENDICES

APPENDIX A: COMPLETE BIBLIOGRAPHIC RECORDS FOR THE 11 RECORDS RETAINED AT THE ELIGIBILITY STAGE

The following records are provided in ABNT format for reference. Records marked with an asterisk received an eligibility verdict at full-text level (record 1); all others were assessed at title and abstract level only.

1. CHEN, M. et al. Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage. ***npj Digital Medicine****, [s. l.], v. 9, art. 409, 2026. DOI: 10.1038/s41746-026-02409-8. [Full text verified. Design: systematic review and meta-analysis. Eligibility verdict: does not meet multi-human team configuration criterion. Individual-level comparisons only.]*

2. DELL'ACQUA, F. *et al.* Navigating the jagged technological frontier: field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. **Organization Science**, [s. l.], 2026. DOI: 10.1287/orsc.2025.21838. [Abstract level only. Design: randomized field experiment. Apparent configuration: individual-level three-condition comparison. Full-text eligibility not confirmed.]

3. DENNIS, A. R.; LAKHIWAL, A.; SACHDEVA, A. AI agents as team members: effects on satisfaction, conflict, trustworthiness, and willingness to work with. **Journal of Management Information Systems**, [s. l.], v. 40, n. 2, 2023. DOI: 10.1080/07421222.2023.2196773. [Abstract level only. Design: 2 × 2 × 2 laboratory factorial experiment. Apparent configuration: AI agent versus human as team member. Full-text eligibility not confirmed.]

4. HAN, J.; REN, R. Unequal AI access enhances team productivity through negative socio-emotional reactions and cognitive diversity. **Proceedings of the Association for Information Science and Technology**, [s. l.], v. 62, n. 1, 2025. DOI: 10.1002/pra2.1437. [Abstract level only. Design: laboratory experiment with two-person teams (n = 60 teams, N = 120 participants). Apparent configuration: three AI access conditions. Full-text eligibility not confirmed.]

5. LIU, Z. *et al.* Who is the best helper for university students' mathematical creativity? A quasi-experimental study of human-ChatGPT, human-Google, and human-human co-creation. **Computer Applications in Engineering Education**, [s. l.], 2025. DOI: 10.1002/cae.70100. [Abstract level only. Design: quasi-experiment with three groups (N =

230). Apparent configuration: human-ChatGPT, human-Google, human-human conditions. Full-text eligibility not confirmed.]

6. PAN, S.; MOU, Y. Team up with AI or human? Investigating candidates' self-categorization as fluidity and ingroup-serving attribution when judged by a human-AI hybrid jury. **International Journal of Human-Computer Interaction**, [s. l.], 2025. DOI: 10.1080/10447318.2024.2408511. [Abstract level only. Design details not extractable from available information. Apparent configuration: hybrid jury conditions. Full-text eligibility not confirmed.]

7. SHAIKH, S. J.; CRUZ, I. F. AI in human teams: effects on technology use, members' interactions, and creative performance under time scarcity. **AI and Society**, [s. l.], v. 38, 2023. DOI: 10.1007/s00146-021-01335-5. [Abstract level only. No extractable design data from available bibliographic information. Full-text eligibility not confirmed.]

8. SHARMA, A. *et al.* Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. **Nature Machine Intelligence**, [s. l.], v. 5, n. 1, p. 46 to 57, 2023. DOI: 10.1038/s42256-022-00593-2. [Abstract level only. Design: randomized controlled trial (N = 300). Apparent configuration: AI-generated feedback intervention in dyadic peer support conversations. Full-text eligibility not confirmed.]

9. YANG, C. *et al.* My advisor, her AI, and me: evidence from a field experiment on human-AI collaboration and investment decisions. **Management Science**, [s. l.], 2026. DOI: 10.1287/mnsc.2022.03918. [Abstract level only. Design: field experiment with additional controlled online experiment. Apparent configuration: pure AI advice versus human-AI collaborative advice. Full-text eligibility not confirmed.]

10. ZHANG, G. *et al.* Trust in an AI versus a human teammate: the effects of teammate identity and performance on human-AI cooperation. **Computers in Human Behavior**, [s. l.], v. 144, art. 107536, 2023. DOI: 10.1016/j.chb.2022.107536. [Abstract level only. Design: 2 × 2 between-subjects experiment. Apparent configuration: AI identity versus human identity × high versus low performance. Full-text eligibility not confirmed.]

11. CHEN, N.; ZHANG, X. Restoring trust in human-centered AI: differential impacts of trust repair strategies in human-AI versus human-human teams. **International Journal of Human-Computer Interaction**, [s. l.], 2026. DOI: 10.1080/10447318.2026.2676762. [Abstract level only. Design: 3 × 2 between-subjects experiment (N = 134). Apparent configuration: three trust repair strategies × human-AI team versus human-human team. Full-text eligibility not confirmed.]