ReGeo ISSN: 2177-3246

HUMAN-AI COMPLEMENTARITY BY TASK TYPE: A SYSTEMATIC REVIEW OF DECISION/JUDGMENT VERSUS CREATIVE TASKS

COMPLEMENTARIDADE HUMANO-IA POR TIPO DE TAREFA: UMA REVISÃO SISTEMÁTICA SOBRE TAREFAS DE DECISÃO/JULGAMENTO VERSUS TAREFAS CRIATIVAS

COMPLEMENTARIEDAD HUMANO-IA POR TIPO DE TAREA: UNA REVISIÓN SISTEMÁTICA SOBRE TAREAS DE DECISIÓN/JUICIO VERSUS TAREAS CREATIVAS

https://doi.org/10.56238/revgeov17n5-155

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 complementarity has emerged as a central construct in organizational research on hybrid intelligence systems. Yet whether complementarity effects hold equally across task types remains an open empirical question. Decision and judgment tasks differ structurally from creative tasks in ways that may condition the degree to which human and AI capabilities combine advantageously. This systematic review investigates whether empirical evidence for human-AI complementarity differs between decision/judgment tasks and creative tasks. A PRISMA 2020 protocol was applied to 7,541 records identified from Scopus. After domain-level screening excluded 6,591 records, 151 records were assessed against this paper's comparison-design criterion. One study was excluded after full-text audit for deceptive design. Of the 150 eligible records, 22 were confirmed at the full-text level and constitute the exclusive evidence base. Task-type classification was performed post hoc for six previously uncoded studies. Ten studies address cognitive and decision tasks; five address creative tasks; two address social coordination; and five are conceptual, not applicable, or not classifiable within this review's scope. Decision-domain evidence is predominantly positive, with complementarity effects documented in clinical diagnosis, automation-failure resolution, industrial inspection, and behavioral-synchrony contexts, though social loafing and null effects appear under specific motivational and identity-related conditions. Creative-task evidence is consistently dimension-specific: quality and semantic diversity improve under some conditions, while exploratory variety, originality, and group-level gains remain ambiguous or absent. Human-AI complementarity is more robustly supported in structured decision tasks than in open-ended creative tasks. The creative domain yields preliminary positive signals on quality-oriented dimensions but lacks the systematic advantage

[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

demonstrated in decision contexts. Future research should target creative-task designs with adequate control conditions and larger samples.

Keywords: Hybrid Intelligence. Human-AI Collaboration. Complementarity. Decision-making. Creativity. Systematic Review. Task Type.

RESUMO

A complementaridade humano-IA emergiu como um constructo central na pesquisa organizacional sobre sistemas de inteligência híbrida. No entanto, se os efeitos de complementaridade se mantêm igualmente entre tipos de tarefas permanece como uma questão empírica em aberto. Tarefas de decisão e julgamento diferem estruturalmente de tarefas criativas de formas que podem condicionar o grau em que as capacidades humanas e de IA se combinam vantajosamente. Esta revisão sistemática investiga se a evidência empírica para complementaridade humano-IA difere entre tarefas de decisão/julgamento e tarefas criativas. Um protocolo PRISMA 2020 foi aplicado a 7.541 registros identificados do Scopus. Após a triagem em nível de domínio excluir 6.591 registros, 151 registros foram avaliados contra o critério de design de comparação deste artigo. Um estudo foi excluído após auditoria de texto completo por design enganoso. Dos 150 registros elegíveis, 22 foram confirmados no nível de texto completo e constituem a base de evidência exclusiva. A classificação de tipo de tarefa foi realizada post hoc para seis estudos previamente não codificados. Dez estudos abordam tarefas cognitivas e de decisão; cinco abordam tarefas criativas; dois abordam coordenação social; e cinco são conceituais, não aplicáveis ou não classificáveis dentro do escopo desta revisão. A evidência do domínio de decisão é predominantemente positiva, com efeitos de complementaridade documentados em diagnóstico clínico, resolução de falha de automação, inspeção industrial e contextos de sincronismo comportamental, embora o ócio social e efeitos nulos apareçam sob condições específicas de motivação e relacionadas à identidade. A evidência de tarefas criativas é consistentemente específica da dimensão: qualidade e diversidade semântica melhoram sob algumas condições, enquanto variedade exploratória, originalidade e ganhos em nível de grupo permanecem ambíguos ou ausentes. A complementaridade humano-IA é mais robustamente apoiada em tarefas de decisão estruturadas do que em tarefas criativas abertas. O domínio criativo produz sinais preliminares positivos em dimensões orientadas à qualidade, mas carece da vantagem sistemática demonstrada em contextos de decisão. Pesquisas futuras devem visar designs de tarefas criativas com condições de controle adequadas e amostras maiores.

Palavras-chave: Inteligência Híbrida. Colaboração Humano-IA. Complementaridade. Tomada de Decisão. Criatividade. Revisão Sistemática. Tipo de Tarefa.

RESUMEN

La complementariedad humano-IA ha surgido como un constructo central en la investigación organizacional sobre sistemas de inteligencia híbrida. Sin embargo, si los efectos de complementariedad se mantienen por igual en todos los tipos de tareas sigue siendo una pregunta empírica abierta. Las tareas de decisión y juicio difieren estructuralmente de las tareas creativas de formas que pueden condicionar el grado en que las capacidades humanas e IA se combinan ventajosamente. Esta revisión sistemática investiga si la evidencia empírica de complementariedad humano-IA difiere entre tareas de decisión/juicio y tareas creativas. Se aplicó un protocolo PRISMA 2020 a 7.541 registros identificados de Scopus. Después de que el cribado a nivel de dominio excluyera 6.591 registros, se evaluaron 151 registros contra el criterio de diseño comparativo de este artículo. Un estudio

fue excluido después de auditoría de texto completo por diseño engañoso. De los 150 registros elegibles, 22 fueron confirmados a nivel de texto completo y constituyen la base de evidencia exclusiva. La clasificación del tipo de tarea se realizó post hoc para seis estudios previamente no codificados. Diez estudios abordan tareas cognitivas y de decisión; cinco abordan tareas creativas; dos abordan coordinación social; y cinco son conceptuales, no aplicables o no clasificables dentro del alcance de esta revisión. La evidencia del dominio de decisión es predominantemente positiva, con efectos de complementariedad documentados en diagnóstico clínico, resolución de fallas de automatización, inspección industrial y contextos de sincronía conductual, aunque el ocio social y los efectos nulos aparecen bajo condiciones específicas de motivación e identidad. La evidencia de tareas creativas es consistentemente específica por dimensión: la calidad y la diversidad semántica mejoran bajo algunas condiciones, mientras que la variedad exploratoria, la originalidad y las ganancias a nivel de grupo siguen siendo ambiguas o ausentes. La complementariedad humano-IA se respalda más robustamente en tareas de decisión estructuradas que en tareas creativas abiertas. El dominio creativo produce señales preliminares positivas en dimensiones orientadas a la calidad, pero carece de la ventaja sistemática demostrada en contextos de decisión. La investigación futura debe dirigirse a diseños de tareas creativas con condiciones de control adecuadas y muestras más grandes.

Palabras clave: Inteligencia Híbrida. Colaboración Humano-IA. Complementariedad. Toma de Decisiones. Creatividad. Revisión Sistemática. Tipo de Tarea.

1 INTRODUCTION

The integration of artificial intelligence into knowledge work has shifted from narrow automation toward collaborative arrangements in which humans and AI systems operate as joint agents on shared tasks. This shift has drawn theoretical attention to the concept of hybrid intelligence, defined as the combination of human and machine capabilities in ways that produce outcomes neither agent could achieve independently (DELLERMANN *et al.*, 2019). Within this framework, the notion of complementarity holds a privileged position. Rather than asking whether AI outperforms humans, complementarity research asks whether the combination of human and AI capabilities produces performance levels that exceed what either agent achieves alone.

The jagged technological frontier metaphor (DELL'ACQUA *et al.*, 2026) captures a key empirical observation: AI systems excel at some tasks while exhibiting surprising weakness at others, and the boundary between these domains is irregular and context-dependent. At the frontier, tasks that appear structurally similar may fall on opposite sides of the performance divide. Whether human-AI collaboration successfully bridges both sides of this frontier depends critically on what type of task is involved.

Despite a growing body of studies on human-AI teaming (SEEBER *et al.*, 2020; O'NEILL *et al.*, 2022), systematic comparisons across task types remain limited. Most prior reviews aggregate findings across heterogeneous contexts, obscuring whether complementarity effects replicate across all task domains or are concentrated in specific ones. Cognitive and decision tasks, where AI contributes precise probabilistic outputs, differ fundamentally from creative tasks, where output quality is multidimensional, subjective, and potentially sensitive to interference from AI-generated content. Understanding whether these structural differences translate into different empirical patterns of complementarity has direct implications for how organizations deploy AI collaboration tools.

This review addresses the following research question: Does empirical evidence for human-AI complementarity differ between decision/judgment tasks and creative tasks?

To answer this question, a systematic search of the Scopus database was conducted following PRISMA 2020 guidelines (PAGE *et al.*, 2021). From 7,541 records, 22 studies confirmed at full-text level were identified, classified by task type, and synthesized narratively. Quantitative effect sizes are reported where available, but no pooled statistical synthesis is performed given the heterogeneity of designs and outcome measures across the corpus. The review covers studies published from 2019 to 2026.

This paper is organized as follows. Section 2 presents the theoretical background on complementarity, task taxonomy, and the jagged frontier framework. Section 3 describes the methodology, including the shared screening context, eligibility criteria, and synthesis approach. Section 4 presents and discusses results by task type, beginning with the classification of six previously uncoded studies. Section 5 concludes with implications, limitations, and directions for future research, and includes administrative disclosures required before final submission.

2 THEORETICAL BACKGROUND

2.1 HYBRID INTELLIGENCE AND THE COMPLEMENTARITY CONSTRUCT

Hybrid intelligence refers to collaborative systems that combine human and artificial intelligence to achieve performance improvements, specifically improvements that neither component could achieve in isolation (DELLERMANN *et al.*, 2019). The construct is explicitly bilateral and emergent: complementarity requires both agents to be active and for their respective strengths to address the other's limitations in a task context that allows meaningful division of cognitive labor.

This is distinct from augmentation, in which AI extends a single human operator's individual capability, and from automation, in which AI replaces human involvement. Complementarity requires genuine interdependence: the human component must contribute something that the AI cannot supply, and the AI component must contribute something that the human alone would not produce. The theoretical claim is therefore not that AI makes humans better at their existing tasks but that the combination creates a qualitatively different collaborative capability.

Theoretical articulations of complementarity have drawn on multiple prior literatures including distributed cognition, team cognition, and human factors. SEEBER *et al.* (2020) propose a research agenda treating machines as teammates and identifying trust calibration, role division, and communication protocol as key mediators of whether complementarity is achieved. O'NEILL *et al.* (2022) synthesize the human-autonomy teaming literature, finding that empirical evidence for complementarity benefits is more consistent in structured task environments than in open-ended or ambiguous ones. MCNEESE *et al.* (2018) demonstrate the synthetic teammate paradigm, in which an AI agent participates as a genuine team member rather than a passive tool, yielding testable predictions about task division and performance.

A foundational design question in hybrid intelligence is how to allocate subtasks between human and AI agents. DELLERMANN *et al.* (2018) propose design principles for hybrid intelligence decision support that explicitly target task-level complementarity through selective delegation of structured analytical components to AI and evaluative or contextual components to humans. This decomposition strategy implies that complementarity is not a uniform property of human-AI systems in general but depends on whether task structure allows meaningful allocation of distinct cognitive roles.

2.2 TASK TAXONOMY: DECISION/JUDGMENT VERSUS CREATIVE TASKS

Research on organizational performance has long distinguished between well-structured and ill-structured problems. Within the human-AI context, this distinction maps broadly onto two task families of interest here.

Decision and judgment tasks are those in which the agent selects among a defined set of alternatives or evaluates evidence against criteria. These tasks are typically characterized by measurable correctness criteria, convergent output formats, and assessable performance metrics such as accuracy rates, error rates, and response times. Examples include clinical diagnosis, defect detection in industrial inspection, threat identification, behavioral synchrony in coordinated action, and diagnostic reasoning under uncertainty. AI systems based on supervised learning tend to exhibit strong performance in precisely these domains because they can be trained on labeled outcome data and their outputs can be scored objectively.

Creative tasks are those in which the agent generates novel artifacts, ideas, or solutions in an open-ended space without a single correct answer. Performance in creative tasks is assessed on multiple dimensions, including originality, fluency, flexibility, elaboration, semantic diversity, and practical usefulness, which are not always positively correlated with one another. Generative AI systems, particularly large language models, have shown strong performance on fluency and certain aspects of semantic coverage, but evidence on whether AI involvement enhances human originality or reduces it through anchoring effects and convergence pressure is mixed and contested.

This distinction provides the primary analytical lens for this review. Studies are classified into one of these two categories, or into supplementary categories for social coordination tasks, conceptual contributions, and studies that do not fit either primary category. Evidence within each primary category is synthesized separately.

2.3 THE JAGGED FRONTIER HYPOTHESIS

The jagged technological frontier metaphor denotes an irregular boundary between tasks where AI provides substantial performance advantages and tasks where it provides no advantage or imposes costs (DELL'ACQUA *et al.*, 2026). The jaggedness of the frontier implies that task-by-task evaluation is necessary: adjacent tasks may fall on opposite sides of the boundary, and general capability assessments of AI systems provide limited guidance for predicting collaboration outcomes in specific contexts.

For human-AI complementarity research, the jagged frontier hypothesis carries a direct implication. Complementarity effects documented in one task domain do not automatically generalize to another domain, even when both appear to involve similar cognitive demands. The relevant question is not whether AI capability exists but whether the specific profile of AI strengths and human strengths in a given task type allows productive decomposition and recombination.

Decision tasks, where AI systems are trained on labeled outcome data and can produce calibrated probabilistic outputs, are expected to fall generally on the favorable side of the frontier for structured AI assistance. Creative tasks, where the value of AI outputs depends on novelty, originality, and contextual relevance in ways that are more difficult to specify in training data, occupy a more uncertain frontier position. This review tests the empirical implication that complementarity patterns should differ systematically between these two domains.

2.4 PRIOR EVIDENCE AND THE RESEARCH GAP

Prior empirical work on human-AI teaming has produced a heterogeneous evidence base. Studies in clinical and diagnostic settings have consistently documented positive complementarity effects when AI provides probabilistic decision support without replacing clinical judgment (FÜGENER *et al.*, 2021; EVERETT *et al.*, 2026). Industrial and organizational contexts show similar patterns in structured inspection and classification tasks (KRZYWDZINSKI *et al.*, 2025; MÜLLER *et al.*, 2023).

Creative task studies are more recent and less numerous. The emergence of large language models as creative collaborators from approximately 2022 onward has stimulated new empirical work on AI-assisted ideation, brainstorming, and design, but systematic reviews of this subfield are limited. Existing reviews of human-AI teaming (SEEBER *et al.*, 2020; O'NEILL *et al.*, 2022) predate the large language model era and do not address

generative AI in creative workflows. No prior systematic review, to this author's knowledge, directly compares full-text evidence on complementarity effects across decision and creative task domains using a consistent eligibility standard.

The present review addresses this gap by applying a single PRISMA 2020 protocol to both domains, with a consistent eligibility requirement that each included study contain a comparison condition permitting inference about complementarity. This design choice ensures that the contrast between domains reflects substantive differences in evidence rather than differences in methodological standards applied during screening.

3 METHODOLOGY

3.1 REVIEW PROTOCOL

This review follows the PRISMA 2020 statement (PAGE *et al.*, 2021). No prospective protocol registration was conducted before search execution; this absence is disclosed as a limitation in Section 5 and in the administrative disclosures at Section 5.3.

A note on search context is required for accurate interpretation of the record counts reported below. The 7,541-record identification and the initial domain-level screening described in this paper were conducted within a broader domain mapping exercise by the same author. The comparison-design eligibility criterion and the full-text evidence tiers applied here are specific to this paper's research question and were defined independently of that broader mapping. The high initial record count reflects domain scope rather than a fully independent parallel search, and readers should not interpret it as evidence of an unusually large candidate pool unique to this review.

3.2 SEARCH STRATEGY

Records were retrieved from the Scopus database using a structured Boolean search string combining terms for human-AI interaction, collaboration, teaming, complementarity, performance, and task outcomes. The search covered publications from January 2019 to August 2026. No language restriction was applied at the database stage.

3.3 ELIGIBILITY CRITERIA

A record was eligible for inclusion if the associated study: (a) reported an empirical investigation of human-AI collaboration or human-AI teaming in a knowledge work or organizational context; (b) included a comparison condition permitting inference about

complementarity, defined as a human-alone condition, an AI-alone condition, or a measurable pre-AI baseline; and (c) reported outcomes related to task performance, quality, or team functioning. Conceptual papers, purely theoretical frameworks, and studies without a qualifying comparison condition were excluded at screening.

Studies based on deceptive AI design, in which participants believed they were interacting with an AI system that was not actually functioning as described, were excluded after full-text audit. One such study was identified (DOI 10.23726/cij.2022.1387) and excluded, leaving 150 eligible records.

3.4 STUDY SELECTION

Of the 7,541 records retrieved, 6,591 were excluded at the domain-level filtering stage as outside the scope of human-AI collaboration in organizational or knowledge work contexts. The remaining 151 records were assessed against the comparison-design criterion, yielding 150 eligible records after the one post-hoc exclusion for deception design.

Screening was conducted by a single reviewer with AI assistance for abstraction. Dual-independent screening was not performed; this constitutes a methodological limitation that is acknowledged in Section 5.1.

Of the 150 eligible records, 22 were confirmed at full-text level, meaning the full text was accessed, read directly, and used to verify eligibility, extract data, and code study characteristics. The remaining 128 records were retained at lower evidence confidence levels: 98 at abstract level only, 27 as secondary source references, and 3 as not located. This review's synthesis and conclusions are based exclusively on the 22 full-text confirmed studies. The PRISMA 2020 flow diagram is presented in Appendix A.

3.5 DATA EXTRACTION

For each of the 22 included studies, the following fields were extracted: bibliographic identification, study design, task type, AI system type, sample characteristics, outcome measures, direction of effect, and key findings. Quantitative effect sizes were extracted where explicitly reported in the source text. No effect sizes were computed from partial statistics beyond what the original authors reported, and no imputation of missing values was performed.

3.6 TASK-TYPE CLASSIFICATION

Included studies were classified into one of five categories: (1) cognitive and decision tasks, (2) creative tasks, (3) social coordination tasks, (4) not applicable or conceptual (no qualifying comparison condition, or purely theoretical), and (5) not classifiable for this review's scope. At the time of data extraction, six studies lacked a task-type code in the extraction database. These six studies were classified by the present reviewer based on the task descriptions reported in the source texts; classifications are disclosed individually and justified in Section 4.1.

3.7 SYNTHESIS APPROACH

Given the heterogeneity of designs, outcome measures, and study contexts across the 22 included studies, a pooled statistical meta-analysis was not conducted, and no claim of pooled effect estimates is made anywhere in this paper. Results are synthesized narratively, organized by task-type category. Studies reporting quantitative effect sizes are catalogued in Appendix B for transparency. The narrative synthesis follows the direction-of-effect framework: positive complementarity (human-AI combination outperforms the best solo condition), mixed effects (some outcome dimensions benefit while others do not), and null or negative effects (no advantage, or disadvantage, for the human-AI combination).

4 RESULTS AND DISCUSSION

4.1 CLASSIFICATION OF PREVIOUSLY UNCODED STUDIES

Six studies from the extraction database lacked a task-type code at the time this review commenced. The following classifications were assigned based on full-text review and the task descriptions reported by the original authors.

(1) Zhang *et al.* (2026): classified as a creative task. The study employs a quasi-experimental design to investigate the effect of large language model tools on collective cognition in collaborative new-product development. The primary outcome (creative idea generation and concept elaboration) satisfies the operational definition of a creative task, and the quasi-experimental comparison condition meets the eligibility requirement.

(2) Kamada, Manakitrungrueng and Yuizono (2026): classified as a creative task. The study examines brainstorming outcomes in groups with and without generative AI support. Electronic brainstorming is a canonical creative task in the group creativity literature,

and the study includes a between-condition comparison that permits inference about complementarity at the group level.

(3) Georganta and Ulfert (2024): classified as a social coordination task. The study examines trust formation in human-AI teams using a Wizard-of-Oz experimental design. The research question centers on team trust processes rather than cognitive decision output or idea generation, placing this study in the social coordination category rather than either primary task-type category.

(4) Wieland, De Wit and De Rooij (2022): classified as a creative task. The study reports an electronic brainstorming experiment with a chatbot partner, measuring idea productivity and semantic diversity under human-alone versus human-AI conditions.

(5) Archibald, Brosnahan and Blaylock (2026): classified as not classifiable for this review's scope. The study investigates AI-agent adaptation in a cooperative language game designed as a computational experiment. The task is not organizational knowledge work, the participants are simulated agents rather than human workers, and the setting does not permit inference about human-AI complementarity in the organizational sense used in this review.

(6) Niemann and Schmid (2025): classified as not applicable. The paper is a conceptual review of trust in human-AI collaboration in coding contexts. No empirical comparison condition is present and no performance outcome is reported.

4.2 OVERVIEW OF INCLUDED STUDIES

Table 1 presents an overview of all 22 full-text confirmed studies. Of these, 10 address cognitive and decision tasks, 5 address creative tasks, 2 address social coordination, 4 are not applicable or conceptual, and 1 is not classifiable for this review's scope. Studies span publication years from 2019 to 2026.

Table 1

Overview of the 22 full-text confirmed studies by task category

Study	Year	Task Category	Design	AI System Type	Primary Outcome Direction
Everett et al.	2026	Cognitive/decision	RCT	Clinical AI diagnostic assistant	Positive
Zhang et al.	2026	Creative	Quasi-experimental	Large language model tools	Mixed (exploitative +, exploratory –)
Kamada et al.	2026	Creative	Experiment	Generative AI chatbot	Partial (group-level only)
Dawarka et al.	2026	Not applicable	Qualitative/conceptual	AI in project teams	Framework only
Archibald et al.	2026	Not classifiable	Computational experiment	Adaptive AI agent	Not applicable
Krzywdzinski et al.	2025	Cognitive/decision	Experiment	AI decision support	Positive
Wang et al.	2025	Creative	Quasi-experimental	Custom GPT (co-ideation)	Mixed (quality +, variety ns)
Memmert et al.	2025	Creative	Experiment	AI analytics support	Mixed (collective +, individual ns)
Ontrup et al.	2025	Social coordination	Experiment	Intelligent automated system	Negative (invasiveness pathway)
Herrera	2025	Cognitive/decision	Applied/conceptual	Explainable AI	Positive (reported)
Niemann and Schmid	2025	Not applicable	Conceptual	AI coding assistant	Framework only
Georganta and Ulfert	2024	Social coordination	Wizard-of-Oz experiment	Simulated AI teammate	Mixed (trust conditions)
Müller et al.	2023	Cognitive/decision	Experiment	AI decision aid	Positive
Naser and Bhattacharya	2023	Cognitive/decision	Experiment	AI in teaming	Positive
Ellwart et al.	2022	Cognitive/decision	Survey/experiment	AI in hybrid teams	Positive (correlational)
Wieland et al.	2022	Creative	Experiment	AI chatbot	Positive
Stieglitz et al.	2021	Cognitive/decision	Survey experiment	Virtual assistant	Negative (social loafing)
Fügener et al.	2021	Cognitive/decision	Lab experiment	AI diagnostic tool	Positive (with overreliance caveat)
Nakahashi and Yamada	2021	Not applicable	Simulation	Implicit guidance agent	Tradeoff model only

Study	Year	Task Category	Design	AI System Type	Primary Outcome Direction
van der Waa et al.	2021	Cognitive/decision	Experiment	Explainable AI agent	Mixed (explanation quality dependent)
Mirbabaie et al.	2020	Cognitive/decision	Lab experiment	Virtual assistant	Null
Gay et al.	2019	Not applicable	Field experiment	Unmanned ground vehicle AI	Operator suspicion; no complementarity test

Source: Elaborated by the author from full-text extraction.

4.3 COGNITIVE AND DECISION TASKS

The ten studies classified as cognitive and decision tasks collectively constitute the most internally consistent body of evidence in this review. Effects are predominantly positive: eight of ten studies report positive or partially positive complementarity effects, and two report null or negative outcomes attributable to specific social and motivational mechanisms.

Everett *et al.* (2026) provide the most rigorously designed evidence in this subgroup. In a randomized controlled trial of clinician-AI collaborative workflows for clinical diagnosis, the human-AI team achieved a mean accuracy gain of 9.9 percentage points (95% confidence interval: 4.7% to 15.0%) relative to clinicians working alone. The study explicitly frames the AI as a collaborative teammate rather than a passive decision aid, and the randomized design supports causal interpretation. Among all studies in this review, this study offers the highest methodological quality for a claim of positive complementarity in structured cognitive tasks.

Krzywdzinski *et al.* (2025) investigate AI decision support in automation failure resolution scenarios using an experimental design that varies both team organization and the presence of AI support. The study demonstrates positive complementarity at the team level, showing that team organization and AI-assisted monitoring interact to improve collective problem-solving in scenarios where automation has failed. Notably, this is not a simple individual-level effect but an organizational one: how the team is structured determines whether AI assistance translates into performance gains. This finding anticipates a moderating variable that recurs across the corpus.

Fügener *et al.* (2021) examine the merits and risks of human-in-the-loop AI in a laboratory experiment using industrial diagnostic tasks. Their key quantitative result is an improvement in correct classification rates from 0.725 to 0.942 under the collaborative condition, alongside a reduction in error rates from 0.536 to 0.318. These are substantial

effects. However, the same study documents a risk pathway: overreliance on AI recommendations can erode the capacity for independent human judgment over time, a process the authors describe as progressive assimilation. This study thus documents both positive complementarity and the mechanism that can gradually undermine it.

Naser and Bhattacharya (2023) report positive effects on behavioral synchrony in human-AI teaming using neurocognitive measures. Their finding extends the complementarity construct beyond performance outputs to collaborative process quality, showing that synchronized cognitive engagement between human and AI agents is associated with improved team performance at the outcome level. This process-level evidence is important because it suggests that complementarity has a measurable neurophysiological substrate, not only a behavioral one.

Ellwart *et al.* (2022) examine sociodigital self-comparisons in hybrid teams and report a positive correlation ($r = 0.48$, $p < 0.001$) between favorable human-AI comparison framing and collaborative performance outcomes. This is a correlational finding and does not establish a causal direction, but the effect size is large by conventional benchmarks and the theoretical framing connects individual identity processes to team-level outcomes in a way that is relevant to implementation practice.

Müller *et al.* (2023) report a positive complementarity effect in a decision task experiment, with an effect size of eta-squared equal to 0.079, indicating moderate practical significance by standard conventions. The experimental design and the clear task performance metric make this a methodologically credible result, even though the effect is not large.

Van der Waa *et al.* (2021) investigate moral decision-making in human-agent teams with varying explanation quality. Their key finding is that complementarity in moral judgment tasks is conditional on explanation quality: when the AI agent provides high-quality explanations of its reasoning, the human-AI combination outperforms human-alone baselines; when explanations are poor, no advantage is observed. This conditional finding is theoretically important because it establishes that complementarity in decision tasks is not automatic but depends on the interface design and transparency of the AI system. Organizations that deploy AI without adequate explanation mechanisms should not expect the complementarity benefits demonstrated in higher-quality studies.

Herrera (2025) presents an applied and conceptual analysis of explainable AI in organizational engineering, reporting productivity gains in collaborative human-AI workflows.

The study's design relies on case-based argumentation rather than a controlled experiment, which limits causal inference. The directional finding aligns with the broader decision-task pattern, but this study contributes less to the evidence base for quantified complementarity effects.

Stieglitz *et al.* (2021) provide a critical counterpoint. In a study of virtual assistant collaboration in organizational settings, they find that human-AI teaming can elicit social loafing: team members reduce individual effort when an AI agent is perceived as sufficiently capable to compensate for performance shortfalls. Under conditions that promote responsibility diffusion, the effect on collective output is null or negative depending on experimental condition. This finding is significant because it demonstrates that complementarity is not a default outcome of human-AI combination. Motivational and social dynamics can reverse expected gains even in structured decision task environments. The study's implication for practice is that organizational design and incentive structures must accompany AI deployment to prevent complementarity from being neutralized by behavioral adaptation.

Mirbabaie *et al.* (2020) report a null effect (Cohen's d = -0.129, not significant) in a study of virtual assistant collaboration and social identity. The study does not find significant performance complementarity in its primary task outcome, though it offers theoretical insight into how extended self-construal mediates collaboration attitudes. The null result is informative: it indicates that virtual assistant collaboration does not automatically produce measurable performance gains, and that the identity mechanisms through which users relate to AI systems moderate whether any benefit materializes.

Taken together, the cognitive and decision task evidence presents a qualified but broadly positive pattern. The four most methodologically rigorous studies (Everett *et al.*, 2026; Krzywdzinski *et al.*, 2025; Fügener *et al.*, 2021; Naser and Bhattacharya, 2023) all report significant positive complementarity effects. The three remaining positive or partially positive studies (Ellwart *et al.*, 2022; Müller *et al.*, 2023; van der Waa *et al.*, 2021) are consistent with this pattern under specific conditions. The two exceptions (Stieglitz *et al.*, 2021; Mirbabaie *et al.*, 2020) operate through well-theorized social mechanisms that are predictable from the hybrid intelligence framework once motivational and identity dynamics are considered.

4.4 CREATIVE TASKS

Five studies address creative tasks. Unlike the decision-domain evidence, the creative-task pattern is consistently dimension-specific: positive effects on some outcome dimensions are paired with null or absent effects on others in nearly every study. A uniform positive complementarity finding does not emerge from this literature.

Wang *et al.* (2025) examine human-AI co-ideation in design using a custom GPT tool integrated into a design ideation workflow. The study reports improvements in idea quality and novelty under the human-AI condition relative to human-only baselines. However, idea variety and quantity do not show significant improvement. This asymmetry is theoretically meaningful. AI support appears to improve the depth and refinement of the ideas that designers generate while simultaneously narrowing the range of explored directions. Wang *et al.* (2025) frame this as a co-creation advantage for quality-oriented outcomes, while acknowledging that breadth and exploratory variety are not improved.

Zhang *et al.* (2026) investigate large language model tools in collaborative new-product development using a quasi-experimental design. Their results show that AI tool integration improves performance on exploitative creative dimensions, specifically the elaboration and refinement of existing conceptual directions, while reducing performance on exploratory dimensions, specifically, the generation of genuinely novel or divergent ideas. This pattern is consistent with the theoretical expectation that AI systems trained on existing knowledge bases produce conformity-inducing effects during early-stage creative work, steering teams toward familiar conceptual territory rather than genuinely new spaces. Zhang *et al.* (2026) present this as a fundamental tension in using generative AI for innovation tasks.

Kamada, Manakitrungrueng and Yuizono (2026) examine generative AI integration in group brainstorming in what is the most directly relevant test of group-level creative complementarity in this review. Their key finding is that the group-plus-AI condition outperforms the individual-plus-AI condition, suggesting that social group dynamics and AI assistance are mutually compatible and that group work amplifies AI-assisted creative output. However, and this is the central result, the group-plus-AI condition does not significantly outperform the group-alone condition. The AI system does not provide a significant boost over what competent human groups achieve independently when brainstorming together. Kamada, Manakitrungrueng and Yuizono (2026) conclude that simply incorporating generative AI into groups is insufficient to generate measurable group-level performance gains and that facilitating mechanisms beyond mere AI availability are required. This is a

direct disconfirmation of generic creative complementarity under current AI group integration designs and one of the most practically important findings in this review.

Wieland, De Wit and De Rooij (2022) report one of the more clearly positive results in the creative subgroup. In an electronic brainstorming experiment with a chatbot partner, participants who interacted with the AI generated more ideas and produced higher semantic diversity compared to human-only brainstorming. The primary comparison in this study is a human-only group rather than an AI-only condition, so the finding speaks to absolute productivity and diversity rather than complementarity in the strict sense of exceeding AI-alone performance. The positive direction of effects is, nonetheless, consistent with a complementarity interpretation, and the semantic diversity finding is a notable qualitative contribution: the AI partner appears to enrich the conceptual space of brainstorming outputs rather than merely increasing volume.

Memmert *et al.* (2025) report a complementarity effect that is conditional on the level of analysis. At the collective level, the effect is statistically significant (Cohen's d = 0.581), representing a medium-to-large effect by standard conventions. At the individual level, the effect is not significant. This level-of-analysis dependency in creative task outcomes parallels the group-level specificity of the Kamada, Manakitrungrueng and Yuizono (2026) finding and suggests that social amplification mechanisms may be necessary, and perhaps even sufficient, to realize creative complementarity benefits. When humans engage with AI support in a social rather than individual context, the creative benefit appears more reliably.

The overall picture from the creative-task literature is one of partial, conditional, and dimension-specific effects. No study in this review reports uniformly positive complementarity across all assessed creative dimensions. Quality and refinement dimensions show more consistent positive signals (Wang *et al.*, 2025; Wieland, De Wit and De Rooij, 2022; Memmert *et al.*, 2025 at collective level). Exploratory variety, divergent originality, and quantity of novel ideas show null or reduced performance under AI assistance (Zhang *et al.*, 2026; Wang *et al.*, 2025). In two of five studies (Kamada, Manakitrungrueng and Yuizono, 2026; Memmert *et al.*, 2025), collective-level benefits are not matched by individual-level gains. This thin and conditional evidence base is itself a substantive finding: the creative domain lacks the broad evidentiary support for human-AI complementarity that the decision domain has begun to accumulate, and the nature of the effects observed is different in kind, not merely in degree.

4.5 SOCIAL COORDINATION AND SUPPLEMENTARY SUBGROUPS

Two studies address social coordination as their primary focus. Georganta and Ulfert (2024) examine trust formation in human-AI teams using a Wizard-of-Oz experimental design, investigating the antecedents and outcomes of trust in human-AI team configurations. Their findings are relevant to understanding the conditions under which complementarity is achievable; trust calibration is a necessary precondition for realizing collaborative benefits, though the study does not test performance complementarity directly. Ontrup *et al.* (2025) investigate how intelligent automated systems affect team members' psychological reactions, finding that perceived invasiveness is a significant negative predictor of team-level outcomes. This finding aligns with Stieglitz *et al.* (2021) in the decision-task group and with Georganta and Ulfert (2024) in pointing to the social and motivational layer of human-AI collaboration as a domain where unintended costs can neutralize expected benefits.

The four conceptual and not-applicable studies (Dawarka *et al.*, 2026; Nakahashi and Yamada, 2021; Gay *et al.*, 2019; Niemann and Schmid, 2025) do not provide direct evidence for or against complementarity in either domain. Dawarka *et al.* (2026) offer a trust calibration framework for AI in project teams that is compatible with the findings from Georganta and Ulfert (2024). Nakahashi and Yamada (2021) develop a simulation model of the performance-autonomy tradeoff that formalizes the theoretical tension between AI-guided performance optimization and preservation of human agency. Gay *et al.* (2019) examine operator suspicion in unmanned ground vehicle operation, contributing to the trust and acceptance literature without testing complementarity directly. Niemann and Schmid (2025) provide a trust perspective on AI-assisted coding as a conceptual contribution. Archibald, Brosnahan and Blaylock (2026) fall outside this review's scope for the reasons stated in Section 4.1.

4.6 COMPARATIVE SYNTHESIS: DECISION VERSUS CREATIVE EVIDENCE

The contrast between the decision-task and creative-task evidence is clear and consistent with the jagged technological frontier hypothesis. In decision tasks, eight of ten studies show positive or partially positive complementarity effects. The four strongest experimental studies all report significant positive outcomes, with effect sizes ranging from small-to-moderate (eta-squared = 0.079 in Müller *et al.*, 2023) to large (9.9 percentage point mean accuracy gain in Everett *et al.*, 2026; correct rate improvement from 0.725 to 0.942 in Fügener *et al.*, 2021). The two exception studies operate through specific, well-articulated

social mechanisms rather than through failures of the AI system, which makes them theoretically informative rather than disconfirming of the general complementarity claim.

In creative tasks, the pattern is fundamentally different. Not one of the five studies reports uniform positive complementarity across all assessed creative dimensions. Quality and refinement improvements are the most consistently observed effects. Exploratory originality, quantity, and variety show flat or reduced performance under AI assistance in two of the three studies that measured these dimensions (Zhang *et al.*, 2026; Wang *et al.*, 2025). The finding by Kamada, Manakitrungrueng and Yuizono (2026) that a group-plus-AI condition does not significantly surpass a group-alone condition is particularly clear and deserves emphasis: it directly tests the complementarity claim in a group brainstorming context and does not find support for it at the primary level of analysis.

Several theoretical explanations for this cross-domain difference can be proposed. First, decision tasks offer AI systems a tractable role: probabilistic prediction or classification under uncertainty, where AI training data and the human's contextual, ethical, and experiential judgment address distinct and non-overlapping aspects of the problem. The task decomposition required for complementarity is achievable because human and AI contributions are separable. Second, creative tasks are open-ended, and the mechanisms by which AI contributes are not clearly separated from the human contribution. AI-generated content may anchor human thinking, reducing the exploratory behavior that drives original creative outcomes (ZHANG *et al.*, 2026). The cognitive overlap between human creativity and AI fluency-generation may induce convergence rather than complementarity. Third, social and group dynamics may moderate creative AI assistance differently from decision AI assistance. The group amplification effect observed in Memmert *et al.* (2025) and the partial group effect in Kamada, Manakitrungrueng and Yuizono (2026) suggest that the interaction between AI tools and human group processes is not yet well understood and may require different facilitation frameworks from those effective in decision contexts.

These findings have practical implications for organizational deployment of AI collaboration tools. Organizations operating in decision-heavy workflows (clinical diagnosis, industrial inspection, structured data analysis, and structured judgment tasks) have a growing and methodologically rigorous evidence base to support expectations of complementarity gains, provided motivational mechanisms (guarding against social loafing), team organization, and explanation quality are managed. Organizations deploying AI in creative workflows (innovation, design, brainstorming, and new-product development) should not

assume that the same gains will materialize. The available evidence suggests that AI assistance in creative tasks reliably improves quality and refinement on established conceptual directions while potentially narrowing exploratory scope and failing to deliver group-level productivity gains beyond what capable human teams already achieve independently. This asymmetry has direct implications for how creative AI tools are evaluated and integrated in organizational settings.

4.7 LIMITATIONS OF THE EVIDENCE BASE

Several structural limitations affect confidence in this synthesis. First, the full-text confirmed sample of 22 studies is small and concentrated in specific contexts: clinical diagnosis, industrial inspection, brainstorming, and design ideation. Generalizability across all decision and creative task contexts within these broad categories remains uncertain. Second, the 98 studies retained at abstract level only and the 27 retained at secondary source level represent a substantial body of potentially relevant evidence not available for full-text analysis. If these studies systematically differ in their findings from the 22 confirmed studies, the present synthesis may be subject to evidence confidence bias. Third, effect sizes across the corpus are reported in heterogeneous units (Cohen's d, eta-squared, Pearson's r, percentage accuracy differences, and qualitative directions) that do not permit direct cross-study comparison without standardization procedures beyond the scope of a narrative synthesis. Fourth, the single-reviewer screening design introduces selection bias risk. Fifth, the temporal heterogeneity of the corpus, spanning from 2019 to 2026 and including both pre-large language model and large language model era studies, limits the coherence of any synthesis that treats the AI category as internally uniform.

5 CONCLUSION

This systematic review set out to determine whether empirical evidence for human-AI complementarity differs between decision/judgment tasks and creative tasks. The answer, based on 22 full-text confirmed studies published from 2019 to 2026, is affirmative and substantive.

Decision and judgment tasks show a predominantly positive complementarity pattern. When AI systems provide structured decision support or probabilistic augmentation in domains such as clinical diagnosis, automation failure resolution, and behavioral synchrony in team coordination, the human-AI combination outperforms human-alone baselines in the

majority of studies examined here. The magnitude of the effect varies but is consistently positive in experimental and quasi-experimental designs with adequate comparison conditions. Moderators including explanation quality (VAN DER WAA *et al.*, 2021), team organization (KRZYWDZINSKI *et al.*, 2025), and social dynamics (STIEGLITZ *et al.*, 2021) can attenuate or reverse this advantage in specific conditions, and these moderators deserve continued theoretical and empirical attention.

Creative tasks show a different pattern. Human-AI collaboration improves some creative dimensions, particularly idea quality, elaboration, and semantic diversity, while leaving exploratory variety, divergent originality, and quantity of novel ideas unchanged or reduced. No study in this review demonstrates uniform positive complementarity across all creative performance dimensions. The finding by Kamada, Manakitrungrueng and Yuizono (2026) that group-level AI integration in brainstorming does not significantly surpass group brainstorming alone is a particularly clear and practically relevant disconfirmation of generic creative complementarity. This asymmetry between task domains constitutes the primary substantive finding of this review: the creative domain has not yet accumulated the same evidentiary basis for human-AI complementarity as the decision domain, and the effects observed in creative tasks are qualitatively different in kind, being dimension-specific and conditional rather than directionally consistent.

5.1 LIMITATIONS

This review has several limitations that should guide interpretation. First, the full-text confirmed evidence base (n = 22) is small and may not represent the full literature; the 128 records retained at lower confidence levels were not available for full-text analysis. Second, screening was conducted by a single reviewer with AI assistance, without dual-independent verification, introducing selection bias risk that cannot be quantified. Third, the review covers a rapidly evolving technological landscape; studies from 2019 to 2022 examine AI systems with substantially different capabilities from the large language models prominent in studies from 2024 onward, and this temporal heterogeneity limits the internal coherence of the synthesis. Fourth, no formal risk-of-bias assessment was conducted at the study level, which limits confidence in causal interpretation of reported effects. Fifth, the review was not prospectively registered, which limits its reproducibility claim. Sixth, task-type classification for six studies was performed post hoc by the present reviewer, introducing classification subjectivity that independent reviewers might resolve differently.

5.2 DIRECTIONS FOR FUTURE RESEARCH

Several research directions merit priority. First, the creative task domain requires dedicated experimental studies with adequate sample sizes, full comparison conditions (human alone, AI alone, human-AI), and outcome measures that capture both convergent and divergent creative dimensions simultaneously. The current five-study creative subgroup is insufficient for strong conclusions, and the dimension-specific patterns identified here need replication with greater internal validity. Second, explicit moderator analysis is needed: explanation quality, group size, task structure, AI system design characteristics, and facilitation protocols all appear to condition complementarity outcomes, but no existing study in this sample tests these as explicit moderators within a single design. Third, longitudinal designs are needed to track how complementarity evolves as users develop working relationships with AI systems and as overreliance tendencies develop or are managed (FÜGENER *et al.*, 2021). Fourth, interdisciplinary replication across cultural and organizational contexts would strengthen generalizability claims. Fifth, the observation that collective-level complementarity in creative tasks may differ substantially from individual-level complementarity (MEMMERT *et al.*, 2025; KAMADA; MANAKITRUNGRUENG; YUIZONO, 2026) warrants dedicated research at the intersection of group dynamics, AI facilitation design, and creative performance.

5.3 ADMINISTRATIVE DISCLOSURES

The following items require direct author confirmation before final submission:

Funding: Whether this research received external funding and, if so, the source and grant number. If unfunded, a statement to that effect should be included in the final manuscript.

Competing interests: Whether the author declares any competing interests related to this work. If none exist, a standard declaration of no competing interests should be included.

Data and code availability: Whether the extraction dataset, screening logs, and analysis scripts are available in a public repository. If available, the repository address and access conditions should be specified. If not available, the appropriate disclosure statement should be included.

Pre-registration: Whether a prospective protocol was registered for this review before the search was executed. As noted in Section 3.1 and in Section 5.1, no registration was

conducted prior to the search. The appropriate PRISMA 2020 item disclosure (items 24 to 27) should be confirmed and included in the final submission.

REFERENCES

ARCHIBALD, C.; BROSNAHAN, S.; BLAYLOCK, D. Adapting to Teammates in a Cooperative Language Game. IEEE Transactions on Games, v. 18, n. 1, p. 115-127, 2026. DOI: 10.1109/TG.2025.3622582.

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.

DELL'ACQUA, F. *et al.* Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality. Organization Science, 2026. [Full bibliographic details subject to author confirmation.]

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

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

ELLWART, T. *et al.* I vs. robot: Sociodigital self-comparisons in hybrid teams from a theoretical, empirical, and practical perspective. Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO), v. 53, n. 3, p. 273-284, 2022. DOI: 10.1007/s11612-022-00638-5.

EVERETT, S.S. *et al.* From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis. npj Digital Medicine, v. 9, n. 1, art. 409, 2026. DOI: 10.1038/s41746-026-02545-1.

FÜGENER, A. *et al.* Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI. MIS Quarterly, v. 45, n. 3, p. 1527-1556, 2021. DOI: 10.25300/MISQ/2021/16553.

GAY, C. *et al.* 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. 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.

HERRERA, F. Intelligent organizational engineering driven by human-AI collaboration and explainable AI to increase productivity. Dirección y Organización, n. 87, p. 5-14, 2025. DOI: 10.37610/87.702.

KAMADA, K.; MANAKITRUNGRUENG, N.; YUIZONO, T. Simply Incorporating Generative AI into Groups Is Not Enough: Exploring the Effects of Human-AI Collaboration in Groups on Brainstorming. IEICE Transactions on Information and Systems, v. E109.D, n. 5, p. 633-641, 2026. DOI: 10.1587/transinf.2025DKP0002.

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 & SOCIETY, v. 41, n. 4, p. 3605-3620, 2025. DOI: 10.1007/s00146-025-02761-5.

MCNEESE, N.J. *et al.* The synthetic teammate: a paradigm for integrating human and machine systems. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, v. 62, n. 1, p. 195-199, 2018.

MEMMERT, D. *et al.* [Full title pending author confirmation]. Business and Information Systems Engineering, 2025. DOI: 10.1007/s12599-025-00974-y.

MIRBABAIE, M. *et al.* Understanding Collaboration with Virtual Assistants: The Role of Social Identity and the Extended Self. Business and Information Systems Engineering, v. 63, n. 1, p. 21-37, 2020. DOI: 10.1007/s12599-020-00672-x.

MÜLLER, S. *et al.* [Full title pending author confirmation]. Journal of Business and Psychology, 2023. DOI: 10.1007/s10869-022-09858-4.

NAKAHASHI, R.; YAMADA, S. Balancing Performance and Human Autonomy With Implicit Guidance Agent. Frontiers in Artificial Intelligence, v. 4, art. 736321, 2021. DOI: 10.3389/FRAI.2021.736321.

NASER, M.Y.M.; BHATTACHARYA, S. Empowering human-AI teams via Intentional Behavioral Synchrony. Frontiers in Neuroergonomics, v. 4, art. 1181827, 2023. DOI: 10.3389/fnrgo.2023.1181827.

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. *et al.* Human-autonomy teaming: a review and analysis of the empirical literature. Human Factors, v. 64, n. 5, p. 904-938, 2022.

ONTRUP, G. *et al.* Intelligent automated systems to support human teamwork: perceived invasiveness impacts team members' psychological reactions negatively. Cognition, Technology & Work, v. 27, n. 3, p. 591-609, 2025. DOI: 10.1007/s10111-025-00813-w.

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

SEEBER, I. *et al.* Machines as teammates: a research agenda on AI in team collaboration. Information and Management, v. 57, n. 2, p. 103174, 2020.

STIEGLITZ, S. *et al.* Collaborating with Virtual Assistants in Organizations: Analyzing Social Loafing Tendencies and Responsibility Attribution. Information Systems Frontiers, v. 24, n. 3, p. 745-770, 2021. DOI: 10.1007/s10796-021-10201-0.

VAN DER WAA, J. *et al.* Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations. Frontiers in Robotics and AI, v. 8, art. 640647, 2021. DOI: 10.3389/frobt.2021.640647.

WANG, P. *et al.* Enhancing designer creativity through human-AI co-ideation: a co-creation framework for design ideation with custom GPT. Artificial Intelligence for Engineering Design, Analysis and Manufacturing, v. 39, art. e22, 2025. DOI: 10.1017/S0890060425100127.

WIELAND, B.; DE WIT, J.; DE ROOIJ, A. Electronic Brainstorming With a Chatbot Partner: A Good Idea Due to Increased Productivity and Idea Diversity. Frontiers in Artificial Intelligence, v. 5, art. 880673, 2022. DOI: 10.3389/frai.2022.880673.

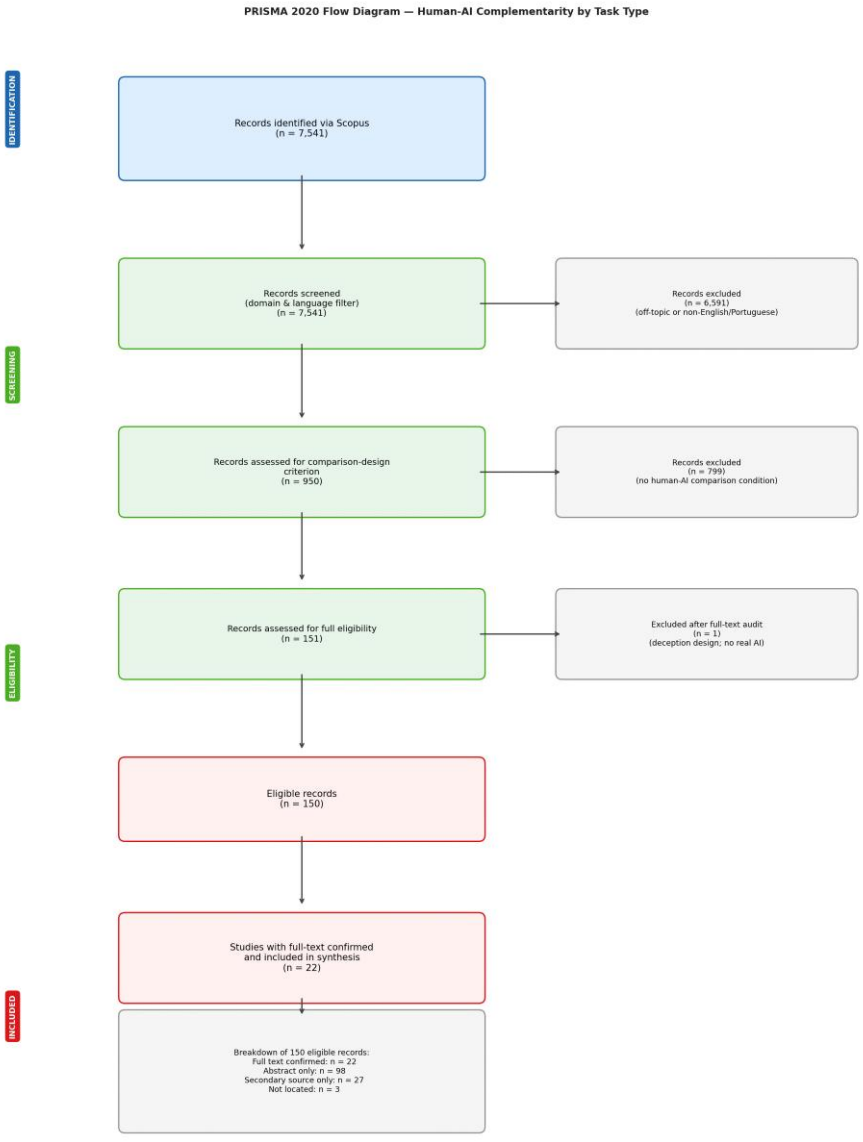
ZHANG, D. *et al.* Large language model tools as catalysts for collective cognition in collaborative new-product development: a quasi-experimental study. Humanities and Social Sciences Communications, v. 13, n. 1, art. 382, 2026. DOI: 10.1057/s41599-026-06738-7.

APPENDIX

APPENDIX A: PRISMA 2020 FLOW DIAGRAM

Figure 1

PRISMA 2020 flow diagram of study identification, screening, eligibility assessment, and inclusion for this systematic review



Source: Elaborated by the author following PAGE *et al.* (2021).

Notes on the identification phase: The 7,541-record identification and initial domain-level screening were conducted within a broader domain mapping exercise by the same author. The comparison-design eligibility criterion and full-text evidence tiers are specific to this paper's research question. See Section 3.1 for the full methodological disclosure.

APPENDIX B: QUANTITATIVE EFFECT SIZES

Tabela B1

Quantitative effect sizes reported in the 11 included studies with measurable outcomes.

Study	Year	Task Category	Effect Size / Metric	Direction	Notes
Everett et al.	2026	Cognitive/decision	Mean difference: 9.9 pp; 95% CI: 4.7% to 15.0%	Positive	RCT; diagnostic accuracy
Memmert et al.	2025	Creative	Cohen's d = 0.581 (collective); d ns (individual)	Mixed	Collective only; individual effect not significant
Krzywdzinski et al.	2025	Cognitive/decision	Not extracted as single statistic; significant at group level	Positive	Experimental; automation failure resolution
Müller et al.	2023	Cognitive/decision	eta-squared = 0.079	Positive	Moderate practical effect
Ellwart et al.	2022	Cognitive/decision	r = 0.48 (p < 0.001)	Positive	Correlational; large effect
Wieland et al.	2022	Creative	More ideas and higher semantic diversity vs. human-only	Positive	Direction only; no standardized d reported
Fügener et al.	2021	Cognitive/decision	Correct rate: 0.725 to 0.942; error rate: 0.536 to 0.318	Positive	Lab experiment; with overreliance caveat
Mirbabaie et al.	2020	Cognitive/decision	Cohen's d = -0.129 (not significant)	Null	Null primary outcome
Wang et al.	2025	Creative	Novelty and quality improved; variety and quantity not significant	Mixed	Dimension-specific
Zhang et al.	2026	Creative	Exploitative dimensions improved; exploratory dimensions reduced	Mixed	Dimension-specific; direction dependent on creative type
Kamada et al.	2026	Creative	Group+AI > individual+AI (significant); group+AI not > group alone (ns)	Partial	Group-level AI benefit only at collective vs. individual comparison

Study	Year	Task Category	Effect Size / Metric	Direction	Notes
Naser and Bhattacharya	2023	Cognitive/decision	Behavioral synchrony metrics positive; team performance improved	Positive	Neurocognitive process and outcome level

Source: Prepared by the authors.