Creativity, Innovation and Entrepreneurship, Vol. 206, 2026, 28–38
https://doi.org/10.54941/ahfe1007380

Understanding How AI-Assisted Evaluation Influences Innovative Processes, Collaboration, and Creative Confidence in Design Education

Jiayue Wang[1], Jiawei Li[2], and Lin Zhu[2]

[1]Beijing Institute of Fashion Technology, No. 2, East Yinghua Road, Chaoyang District, Beijing 100029, China
[2]Academy of Art and Design, Tsinghua University, Beijing 100084, China

ABSTRACT

Generative AI is increasingly embedded in design practice, yet its effects on convergent stages such as concept screening and selection remain underexplored in team settings. This study examines how a transparent AI co evaluator shapes screening practices, collaboration, and creative confidence in design education. We conducted a mixed-methods workshop where design student teams ranked ideas, then reconsidered choices after AI feedback. ProSight is a co-evaluator that returns rubric scores with brief evidence-linked rationales from each concept's image and description. Teams reviewed the feedback, discussed agreement and conflict, and finalized selections. Data sources included ranking trajectories, pre post surveys on trust in AI, teamwork collaboration, and creative confidence, as well as post task discussions and interaction logs. Across groups, AI input rarely replaced initial preferences, but it shifted how options were compared and justified by making evaluative criteria more explicit and surfacing considerations that were less salient in early deliberation. Participants treated the AI as a credible reference for reflection and justification, but verified suggestions against intent and constraints, and kept authority over trade-offs and final decisions. The findings clarify how explanation rich AI can support reflective screening while reshaping collaborative dynamics in time limited group decision making.

Keywords: Human–AI collaboration, Transparent AI, Team decision-making, Design education, Creativity support tools

INTRODUCTION

The increasing use of generative AI in design is transforming the way decisions are made at every stage of the process. Prior research has largely concentrated on divergence-oriented support such as ideation and inspiration, including work showing how LLMs can be embedded into group ideation workflows and extended to support evaluation-related activities (Shaer et al., 2024). What remains less understood is how AI reshapes the convergent moments of screening, ranking, and selection, where ideas either move forward or quietly drop out. Those moments are pivotal for innovation because teams

Received February 15, 2026; Revised April 5, 2026; Accepted April 20, 2026; Available online July 20, 2026

must converge under time pressure, with criteria that are often incomplete, evolving, or contested.

This gap matters because adding AI does not automatically translate into better team decisions. A meta-analysis across human-AI experiments reports a consistent pattern: human–AI combinations tend to yield clearer gains in content-creation tasks, while decision-making tasks show smaller gains and can underperform the best individual decision maker (Vaccaro et al. 2024). In design teams, evaluation is not only a scoring act; it is a negotiated practice shaped by shared criteria, disagreement handling, and shifting authority. When AI enters this negotiation, the questions go beyond decision quality and extend to how teams reason together, how conflict is resolved, and whether creative confidence is supported or quietly eroded.

Transparency and explainability are commonly proposed as guardrails, yet the evidence base is still uneven for collaborative evaluation settings. Human-centered work on "LLM-as-a-judge" emphasizes keeping humans involved in setting criteria and interpreting judgments, rather than treating model outputs as a drop-in replacement for evaluation (Pan et al., 2024a). In parallel, explainable AI research shows that explanations can improve trust calibration and task performance when they help people verify or contest AI outputs (Naiseh et al., 2023a; Senoner et al., 2024). Socially situated accounts also remind us that explanations operate inside interactional and organizational contexts, where they can reshape accountability, attention, and who gets to "own" the decision (Ehsan et al., 2021). For collaborative design screening in particular, there is still limited empirical insight into how transparent AI feedback is read, debated, and integrated, and what that does to collaboration quality, creative confidence, and perceived agency over the course of the work.

To address these gaps, this study examines AI-assisted concept screening as an emerging design practice. We introduce ProSight, a transparent AI co-evaluator, into team-based screening to study how it influences evaluative reasoning, collaboration dynamics, and creative agency during ranking and selection. The investigation is guided by the following research questions:

RQ1: How does AI-assisted evaluation influence team-based concept screening and selection practices?

RQ2: How does transparent AI-assisted evaluation differentially affect collaboration experience, trust in AI, and perceived creative confidence during design screening?

AI-Assisted Evaluation in Convergent Design: Decision Quality, Justification, and Novelty

In recent years, artificial intelligence (AI) and large language models (LLMs) have been increasingly integrated into design workflows, particularly in convergent stages such as concept screening, selection, and critique. Empirical work in this space points to a fairly consistent takeaway: AI and LLM-assisted evaluation can improve decision quality in convergent tasks, often through faster comparisons and more stable judgments, while also nudging teams toward more explicit justifications and, in some settings, widening the

pool of candidates considered. These effects are conditional and vary with task framing, domain constraints, and evaluation protocols.

Across architecture, engineering and material selection, and design pedagogy, studies report moderate-to-high agreement between LLM outputs and expert assessments, in some cases approaching expert-like performance. This makes LLM-supported screening a plausible "second opinion" when attention is scarce or time is tight (Stige et al., 2023; Yüksel et al., 2023). Beyond speed, several studies note that LLM-based evaluators often provide more clearly structured rationales. They tend to lay out criteria explicitly, summarize trade-offs more cleanly, and use consistent comparison wording. This structure can make evaluation records easier to follow, contest, and document in group settings (Medina & Murakami, 2025). A recurring limitation also appears: rationales may read coherent and well organized but remain weakly grounded in project-specific constraints or tacit domain knowledge, particularly when inputs are underspecified or when prompts cannot capture the nuance of expert practice (Neema et al., 2025; Soliman & Keim, 2025).

On innovation and novelty, experiments on creativity evaluation suggest that LLM outputs can correlate with human novelty ratings and occasionally help surface options perceived as more original. Yet multiple studies echo a practical tension: novelty can coincide with thinner feasibility checks, less elaboration, or vague operational detail, which argues for using AI to broaden candidate sets rather than to close decisions on its own (Chandrasekera et al., 2024; Nowak et al., 2025; Zhang et al., 2025). A parallel line of work documents limitations in LLM judging, including a tendency to reinforce prior choices, sensitivity to prompting, and variability across models and task types. Taken together, AI and LLM-assisted evaluation shows promise for improving decision quality and justificatory reasoning, and sometimes for elevating novelty, but gaps remain. Relatively few in-situ team studies connect AI-assisted screening to process dynamics, such as how criteria are negotiated, how disagreements are handled, and how confidence shifts. Evidence is also uneven in complex interdisciplinary projects, and causal accounts remain incomplete regarding how prompting and collaboration protocols shape the balance among novelty, feasibility, and justification quality.

Human–AI Transparency in Creative Collaboration: Mechanisms Shaping Decision Dynamics, Collaboration Quality, and Agency

In creative teamwork, the practical value of AI transparency lies less in "making the model right" than in how explanations shape agreement, productive disagreement, and the distribution of agency during evaluation. Research on human-AI collaboration shows a mixed pattern: human-AI combinations can help with exploratory or generative work, but in decision-focused settings they may fail to outperform the best individual, especially when AI advice is treated as authority rather than material for negotiation (Choudhary et al., 2023; Shi et al., 2023; Vaccaro et al., 2024a). In this sense, alignment and divergence are not just performance markers; they function as process variables that shape how teams compare options and justify decisions. Studies describe recurring

behaviors once AI recommendations appear, including selective adoption, confirmation seeking, second-look reviews, and criteria recalibration. These behaviors can strengthen decisions when they trigger reflective checks, yet they can also create friction when disagreement persists or when teams lack a workable protocol for integrating AI input (Ding et al., 2025; Hauptman et al., 2024; Steyvers & Kumar, 2023; Vinchon et al., 2023).

A central argument in XAI is that transparency supports trust calibration and shared understanding. Evidence from high-stakes domains suggests that explanation formats, when aligned with the task and readable to users, can improve human-AI team performance and help people calibrate reliance to system capability (Morrison et al., 2023; Silva et al., 2022; Zercher et al., 2025). Explanations are also framed as accountability artifacts that teams can reference when auditing decisions or articulating "why we chose this," which matters in collaborative settings where responsibility is distributed (Endsley, 2022). In practice, these benefits tend to show up when explanations become a common reference that anchors discussion, reduces impression-led debate, and helps teams make criteria speakable.

Transparency, though, comes with costs and can reshape interaction in unhelpful ways. Explanation-rich systems may raise cognitive load and coordination effort, particularly under time pressure, and can shift conversational authority toward AI outputs by presenting a seemingly formal rationale that feels harder to challenge (Hauptman et al., 2024). Work in innovation contexts also suggests that highly transparent systems may dampen creative risk-taking for some participants, especially those with lower self-efficacy, by encouraging reliance on well-justified options (Jiang et al., 2023). The emerging picture is uneven: transparent AI can support sensemaking and accountability while also reconfiguring collaboration quality and agency across team members. A key gap is the limited number of design-relevant, in-situ studies that trace these mechanisms during convergent screening and selection, linking explanation use to shifts in interaction patterns, creative confidence, and collaborative agency during real evaluation work.

To address these gaps, we developed ProSight, a transparent AI co-evaluator that offers rubric-grounded feedback as a reference during team screening. The tool is designed to support comparison and justification, helping teams surface criteria, and reach final selections through their own negotiation.

AI Scoring Agent: Architecture and Design

ProSight is an AI scoring agent built on Coze to support structured human-AI co-evaluation during open-ended design screening. It uses a multimodal, retrieval-supported pipeline: teams submit a prototype image with an optional short description, and ProSight returns rubric-aligned scores with concise rationales grounded in retrieved references. The system is designed to keep the evaluation procedure stable, the reasoning traceable to internal knowledge resources, and the output format consistent for cross-team comparison.

The workflow has three modules: (1) input processing, which extracts key intent and constraints from text and salient interaction cues from images; (2) retrieval-supported evaluation, which queries two internal resources:

an item-definition repository encoding rubric semantics and lightweight decision rules, and a case reference library of annotated concept examples with quality-band summaries and reviewer comments; and (3) feedback generation, which composes criterion-level explanations tied to rubric meanings and comparable cases. When text is missing, ProSight relies more on image cues and retrieved definitions and exemplars.

ProSight embeds a rubric adapted from our prior framework on foresight prototyping in HCI, which operationalizes how speculative prototypes support imagination, reflection, and anticipation through four perceptual dimensions: **Future Adaptability, Functional Visibility, Sensory Experience**, and **Creative Divergence**. These dimensions also structure the internal case library used for retrieval. We add **Aesthetic Quality** as a supplementary dimension to make presentation-related judgments explicit during screening.

Given extracted cues and retrieved references, ProSight outputs a five-dimensional score vector plus an overall score, followed by dimension-level rationales that point to the supporting evidence. It also generates brief rubric-aligned improvement suggestions indicating where evidence is weak and how the concept could be strengthened. To keep feedback comparable across submissions, the agent uses a fixed rubric, a fixed output template, and a stable retrieval configuration. The system exports structured interaction logs (inputs, outputs, session IDs, latency, and token usage) to support analysis of efficiency, feedback characteristics, and adoption.

METHODS

The study was conducted in a design education workshop on service design for space mission environments, designed to elicit forward-looking problem framing and concept justification beyond surface aesthetics. Thirty undergraduate design students (ages 19 to 22) participated voluntarily and worked in six teams.

The procedure had four phases: (1) orientation and rubric familiarization, where teams received the brief, deliverables, and rubric exemplars to establish a shared evaluative vocabulary; (2) pre-AI screening, where teams generated concepts, shortlisted candidates, and produced an initial rank order using human judgment, documenting the initial rank, key criteria with a one-sentence rationale, and one unresolved trade-off; (3) AI-assisted evaluation, where teams submitted each shortlisted concept to ProSight via a prototype image and optional description, reviewed the output for every submission, and recorded the most relevant AI points, their stance (accept, partially accept, reject) with justification, and any rank revision, citing at least one AI statement; and (4) post-AI selection and reflection, where teams finalized rankings, completed a brief team reflection, and then an individual post-task questionnaire.

Agent role and prompt specification. ProSight was configured as a rubric-based co-evaluator. The prompt defined the agent as an industrial design analyst and specified inputs (prototype image, optional concept text, retrieved rubric semantics and reference cases). The role instruction stated:

"*As a professional and experienced industrial design analyst… [to] provide structured, rubric-aligned co-evaluation… and produce transparent rationales that can be audited by human evaluators*."

The task required (i) an objective description of the prototype, (ii) extraction of key design features, (iii) scoring on five rubric dimensions using a 1–5 scale with evidence-based rationales, and (iv) concise improvement suggestions. Output followed a fixed section template to support comparison across concepts.

The study combined self-report measures, behavioral indicators, and qualitative materials. Pre-post questionnaires assessed trust in the AI tool, perceived collaboration quality, and creative confidence. Behavioral indicators captured rank-order alignment between human and AI evaluations (Spearman's rho) and changes in concept trajectories from pre-AI to final rankings. Qualitative data included team decision logs, post-task discussion audio, and exported agent transcripts, which were analyzed to characterize how AI feedback was referenced, contested, and incorporated during screening.

RESULTS

To assess alignment between AI-assisted scoring and teams' initial judgments, we computed Spearman's rank correlations between each team's pre-screening rankings and ProSight's rankings, using rho because teams ranked only three to four concepts and the data were ordinal. Correlations were generally positive across teams, indicating that AI recommendations tended to track human priorities while still producing local divergence (Table 1). Given the small within-team item counts, only one team reached statistical significance, (Group 4: $\rho = 1.00$, $p = 0.000$), which is unsurprising given the small within-team item counts. When pooling concepts across all teams, the association was moderate and statistically significant, indicating systematic alignment overall without implying redundancy.

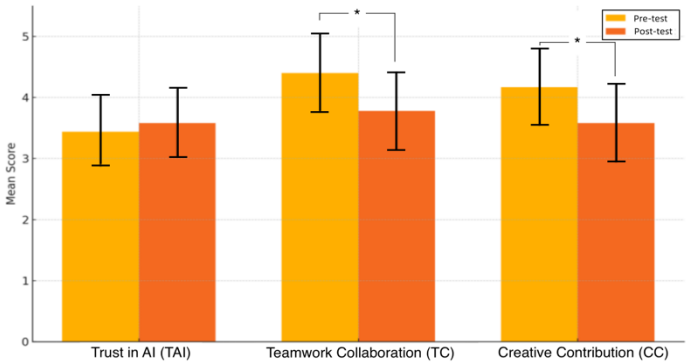
Table 1: Human–AI ranking consistency across groups.

Group NO.	Number of Concepts (n)	Spearman's ρ	p-value
Group 1	4	0.8	0.200
Group 2	3	0.5	0.667
Group 3	4	0.4	0.600
Group 4	3	1	<0.001
Group 5	4	0.2	0.800
Group 6	4	0.6	0.400
Overall	22	0.58	0.006

Note: Spearman's rank correlation (ρ) was used due to the ordinal nature of ranking data and the small number of concepts per group.

Final selections were typically drawn from concepts the agent ranked near the top, while occasional lower-ranked choices indicate that teams treated AI output as a comparative reference rather than a binding rule.

To evaluate changes in participants' perceptions before and after using the AI scoring system, paired-samples t-tests were conducted on three dimensions: Trust in AI (TAI), Teamwork Collaboration (TC), and Creative Confidence (CC). The key results are summarized below in Figure 1.



*: $p < 0.05$

Figure 1: Pre–post comparison of self-reported trust, teamwork collaboration, and creative confidence.

Trust showed a small, non-significant increase from pre to post (M = 3.44 to 3.58), indicating that brief exposure did not meaningfully shift overall reliance attitudes. In contrast, both collaboration experience and creative confidence declined substantially after the AI-assisted phase. Team collaboration dropped from 4.40 to 3.78, $p < 0.001$ and creative confidence decreased from 4.17 to 3.58, $p < 0.001$. Together, these results suggest a short-term pattern in which the tool did not change trust, yet coincided with reduced perceived collaboration quality and lower confidence in creative confidence, consistent with the possibility that explanation-driven comparison increased coordination demands during time-limited screening.

Following the AI-assisted evaluation, each group completed an immediate post-task discussion to capture how they interpreted the AI feedback and how it affected screening. Discussions followed a short semi-structured guide focused on trust in the AI evaluation, perceived influence on prioritization, and comparisons with team judgment; sessions were audio-recorded with consent, transcribed verbatim, and analyzed using thematic analysis with a hybrid coding approach.

Participants described the agent as credible but not decisive. They valued its perceived neutrality and systematic coverage, often using it as a second opinion to surface criteria they had overlooked and to prompt a second review that helped sharpen justifications. At the same time, reliance was cautious: teams reported checking AI suggestions against their intent and constraints, and expressed skepticism when feedback felt vague or poorly aligned. In line with this calibrated stance, AI input seldom changed final priorities. It more often confirmed earlier choices and strengthened confidence by giving language and structure to concerns already present. Participants repeatedly

framed the interaction as complementary: the AI expanded the evaluative lens, while humans retained authority for contextual trade-offs and final decisions.

Discussion: AI Support for Reflective Screening in Design Education

Across teams, AI and human rankings were moderately aligned and final choices rarely shifted, yet teams' justifications and discussion patterns changed after AI input. This supports a view of AI assisted screening as a reference that prompts comparison, surfaces overlooked criteria, and helps teams articulate trade-offs, while keeping selection authority with the group. The pattern is consistent with evidence that human AI combinations often do not outperform the best solo decision maker in decision focused tasks, even when AI adds structure to reasoning (Vaccaro et al., 2024). It also aligns with human centered guidance for using LLMs as evaluators, which emphasizes keeping criteria aligned with human intent and treating model judgments as configurable, inspectable inputs to deliberation.

A plausible mechanism is that explanation driven feedback increased the salience of criteria and exposed gaps in teams' earlier reasoning, which improved justification but also raised coordination demands during time limited discussion. Prior work suggests that explanations can increase acceptance of AI recommendations and do not automatically yield complementary team performance, which fits the idea that transparency reshapes interaction even when outcomes remain stable (Bansal et al., 2020). The stable trust scores alongside drops in perceived teamwork collaboration and creative confidence are also compatible with trust calibration accounts where explanation form and interface fit matter for reliance, and with findings that interaction designs can reduce overreliance by forcing more reflective processing (Naiseh et al., 2023). These results point to practical implications for design education tools: prioritize readable criterion based rationales that support team debate, and add interaction cues that encourage scrutiny instead of steering groups toward a single "correct" ranking (Pan et al., 2024).

Several limitations temper the interpretation. The workshop involved a small sample, few concepts per team, and short exposure, which constrains statistical power and makes effects sensitive to task timing and group composition. In addition, feedback quality depended on the fidelity of submitted images and descriptions and on the coverage of internal knowledge resources, so generalization to other design domains and longer projects remains uncertain.

CONCLUSION

This study suggests that AI-assisted evaluation in design education shifts how teams screen ideas more than what they ultimately select. Across teams, AI rankings showed moderate agreement with human pre-screening, and final choices were usually stable; the main contribution of the agent appeared in the conversation around decisions. AI feedback helped teams compare options more explicitly, notice criteria they had downplayed, and produce clearer justifications. At the same time, transparency came with short-term

trade-offs: trust remained largely unchanged, while perceived collaboration quality and creative confidence declined, pointing to added coordination load and subtle changes in how attention and authority were distributed during time-limited discussion. These findings frame transparent AI tools as collaborative scaffolds that support reflective screening without displacing peer negotiation.

Future studies should vary briefing, interaction protocols, and explanation presentation to identify configurations that maintain interpretability while reducing coordination strain. Longer deployments across different design tasks and cohorts are also needed to assess whether these short-term effects persist, fade, or invert as teams gain experience with AI-assisted screening.

ACKNOWLEDGMENTS

Project supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 52405279) and Beijing Digital Education Research General Project (Grant No.BDEC2025619131), and Fundamental Research Funds for the Municipal Universities of Beijing (KJCX251906).

REFERENCES

Bansal, G., Wu, T. S., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2020). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445717

Chandrasekera, T., Hosseini, Z., & Perera, U. (2024). Can artificial intelligence support creativity in early design processes? *International Journal of Architectural Computing*, *23*, 122–136. https://doi.org/10.1177/14780771241254637

Choudhary, V., Marchetti, A., Shrestha, Y., & Puranam, P. (2023). Human-AI Ensembles: When Can They Work? *Journal of Management*, *51*, 536–569. https://doi.org/10.1177/01492063231194968

Ding, S., Pan, X., Hu, L., & Liu, L. (2025). A new model for calculating human trust behavior during human-AI collaboration in multiple decision-making tasks: A Bayesian approach. *Comput. Ind. Eng.*, *200*, 110872. https://doi.org/10.1016/j.cie.2025.110872

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). *Expanding explainability: Towards social transparency in AI systems*. 1–19. https://doi.org/10.1145/3411764.3445188

Endsley, M. (2022). Supporting Human-AI Teams: Transparency, explainability, and situation awareness. *Comput. Hum. Behav.*, *140*, 107574. https://doi.org/10.1016/j.chb.2022.107574

Hauptman, A., Schelble, B., Duan, W., Flathmann, C., & Mcneese, N. (2024). Understanding the influence of AI autonomy on AI explainability levels in human-AI teams using a mixed methods approach. *Cognition, Technology & Work*, *26*, 435–455. https://doi.org/10.1007/s10111-024-00765-7

Jiang, J., Karran, A. J., Coursaris, C. K., Léger, P.-M., & Beringer, J. (2023). A Situation Awareness Perspective on Human-AI Interaction: Tensions and Opportunities. *International Journal of Human–Computer Interaction*, *39* (9), 1789–1806. https://doi.org/10.1080/10447318.2022.2093863

Medina, I. F. V., & Murakami, T. (2025). An Explainable Artificial Intelligence Framework for Leveraging Large Language Models in Early-Stage Product Design. *Volume 4: 22nd International Conference on Design Education (DEC); 30th Design for Manufacturing and the Life Cycle Conference (DFMLC); 37th International Conference on Design Theory and Methodology (DTM)*. https://doi.org/10.1115/detc2025-164061

Morrison, K., Spitzer, P., Turri, V., Feng, M., Kuhl, N., & Perer, A. (2023). The Impact of Imperfect XAI on Human-AI Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, *8*, 1–39. https://doi.org/10.1145/3641022

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, *169*, 102941. https://doi.org/10.1016/j.ijhcs.2022.102941

Neema, S., Jha, S., Nagel, A., Lew, E., Sureshkumar, C., Gordic, A., Shimmin, C., Nguygen, H., & Eremenko, P. (2025). On the Evaluation of Engineering Artificial General Intelligence. *ArXiv*, *abs/2505.10653*. https://doi.org/10.48550/arxiv.2505.10653

Nowak, R., Figge, P., & Haeussler, C. (2025). AI-Based Measurement of Innovation: Mapping Expert Insight into Large Language Model Applications. *ArXiv*, *abs/2508.02430*. https://doi.org/10.48550/arxiv.2508.02430

Pan, Q., Ashktorab, Z., Desmond, M., Santillán Cooper, M., Johnson, J., Nair, R., Daly, E., & Geyer, W. (2024). *Human-centered design recommendations for LLM-as-a-judge* (N. Soni, L. Flek, A. Sharma, D. Yang, S. Hooker, & H. A. Schwartz, Eds; pp. 16–29). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.hucllm-1.2

Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., & Netland, T. (2024). Explainable AI improves task performance in human–AI collaboration. *Scientific Reports*, *14*. https://doi.org/10.1038/s41598-024-82501-9

Shaer, O., Cooper, A., Mokryn, O., Kun, A. L., & Ben Shoshan, H. (2024). *AI-augmented brainwriting: Investigating the use of LLMs in group ideation*. 1–17. https://doi.org/10.1145/3613904.3642414

Shi, Y., Gao, T., Jiao, X., & Cao, N. (2023). Understanding Design Collaboration Between Designers and Artificial Intelligence: A Systematic Literature Review. *Proceedings of the ACM on Human-Computer Interaction*, *7*, 1–35. https://doi.org/10.1145/3610217

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., & Gombolay, M. (2022). Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. *International Journal of Human–Computer Interaction*, *39*, 1390–1404. https://doi.org/10.1080/10447318.2022.2101698

Soliman, M., & Keim, J. (2025). Do Large Language Models Contain Software Architectural Knowledge? : An Exploratory Case Study with GPT. *2025 IEEE 22nd International Conference on Software Architecture (ICSA)*, 13–24. https://doi.org/10.1109/icsa65012.2025.00012

Steyvers, M., & Kumar, A. (2023). Three Challenges for AI-Assisted Decision-Making. *Perspectives on Psychological Science*, *19*, 722–734. https://doi.org/10.1177/17456916231181102

Stige, Å., Zamani, E., Mikalef, P., & Zhu, Y. (2023). Artificial intelligence (AI) for user experience (UX) design: A systematic literature review and future research agenda. *Inf. Technol. People*, *37*, 2324–2352. https://doi.org/10.1108/itp-07-2022-0519

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, *8*, 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Vinchon, F., Lubart, T., Bartolotta, S., Gironnay, V., Botella, M., Bourgeois-Bougrine, S., Burkhardt, J.-M., Bonnardel, N., Corazza, G., Glăveanu, V., Hanson, M. H., Ivcevic, Z., Karwowski, M., Kaufman, J., Okada, T., Reiter-Palmon, R., & Gaggioli, A. (2023). Artificial Intelligence & Creativity: A Manifesto for Collaboration. *The Journal of Creative Behavior*. https://doi.org/10.1002/jocb.597

Yüksel, N., Börklü, H., Sezer, H., & Canyurt, O. (2023). Review of artificial intelligence applications in engineering design perspective. *Eng. Appl. Artif. Intell.*, *118*, 105697. https://doi.org/10.1016/j.engappai.2022.105697

Zercher, D., Jussupow, E., Benke, I., & Heinzl, A. (2025). How Can Teams Benefit From AI Team Members? Exploring the Effect of Generative AI on Decision-Making Processes and Decision Quality in Team–AI Collaboration. *Journal of Organizational Behavior*. https://doi.org/10.1002/job.2898

Zhang, J., Han, J., & Ahmed-Kristensen, S. (2025). Exploring the use of LLMs to evaluate design creativity. *Proceedings of the Design Society*, *5*, 1773–1782. https://doi.org/10.1017/pds.2025.10191