Toward User-Orchestrated Human–Multi-Agent Teams for Co-Creation

Hyunseung Lim [1] **Hwajung Hong** [1]

Abstract

The use of technology to support human creativity now extends to multi-agent systems (MAS), in which multiple specialized agents contribute distinct functions to a shared creative process. MAS offer new possibilities for co-creation, yet many current systems remain oriented toward autonomous task execution, with humans positioned as prompters or evaluators rather than active members of the creative team. In this short position paper, we argue that creative co-creation calls instead for user-orchestrated Human–Multi-Agent Teams (HMATs), teams in which humans participate alongside multiple agents while retaining the ability to keep the team aligned with the evolving purposes of the work. Drawing on our previous empirical study of how design practitioners form and operate such teams, we develop this position and sketch a three-layer research agenda spanning team formation, interaction design, and agent capabilities. More broadly, we suggest that progress on user-orchestrated HMATs will require both AI/ML research dedicated to these challenges and sustained collaboration between the AI/ML and HCI communities.

1. Introduction

Computational creativity has long explored how technology can support and enhance the production of creative artifacts, ideas, and performances (Colton et al., 2012; Frich et al., 2021). A central focus within this tradition has been the development of creative support tools (CSTs), which aid engagement in inherently creative tasks (Shneiderman, 2002). Recent advances in generative AI are now reshaping the role and mode of CSTs by endowing AI systems with capabilities for open-ended generation, natural-language interaction, and contextual reasoning (Rezwana & Maher, 2023). These agents take on cognitively distinct roles that mirror the complementary functions of real-world creative teams, ranging from productive partners that generate content to reflective guides that critique and steer (Xu et al., 2025; Khan et al., 2025). The rise of such role-specialized agents is shifting creative workflows from single-agent interaction toward multi-agent systems (MAS), in which multiple agents contribute distinct functions to a shared creative process. For example, DesignGPT assigns product manager, design director, and CMF designer agents to different dimensions of product design, enabling agents to contribute distinct perspectives across the design process (Ding et al., 2023). This configuration positions MAS as a new form of creative support tool, an emerging paradigm spanning a growing range of creative domains (Lin et al., 2025).

Despite the promise of generative AI-powered MAS for co-creation, many current MAS implementations remain oriented toward automation, distributing creative roles across multiple agents while offering limited opportunities for sustained human participation (Dahiya et al., 2023; Lin et al., 2025). For example, prior work has raised concerns that assigning idea generation roles to AI agents can introduce or exacerbate design fixation, narrowing the solution space and leading to less original outcomes (Wadinambiarachchi et al., 2024; E et al., 2024). Moreover, delegating creative production to agents may disrupt designers' creative thinking processes or undermine the development of their creative capabilities (Shin et al., 2025). Together, these concerns underscore that developing MAS for co-creation should be approached with particular care to avoid inadvertently constraining users' creativity and innovation. Despite growing recognition of these risks, we still lack a clear understanding of how MAS can be designed to harness the co-creative potential of multiple agents while sustaining humans' active participation in the creative process.

In this position paper, we argue that creative co-creation requires moving beyond fully autonomous Multi-Agent Systems toward user-orchestrated Human–Multi-Agent Teams (HMATs), a configuration we view as an under-operationalized problem setting for ML research, where the question is not whether agents can complete tasks autonomously but how reliably they can be directed by a user, retain that direction, and stay aligned with the user's evolving creative intent over time. This problem setting has re-

[1]Department of Industrial Design, KAIST, Daejeon, Republic of Korea. Correspondence to: Hwajung Hong <hwajung@kaist.ac.kr>.

ICML'26 Workshop on Human-AI Co-Creativity, Seoul, South Korea.

ceived limited attention in the ML literature, in part because articulating it draws on a vantage that is more developed in HCI, on how creative teams actually form and operate. One contribution of this paper is to translate that vantage into an agenda that the ML community can take up. We offer this as a synthesis that recasts existing work around an orchestrating user, grounded in an illustrative case study rather than definitive evidence, and intended to open this problem setting rather than to settle it.

This claim involves two reframings. First, MAS for co-creation should not treat humans as external clients who supply prompts and receive outputs, but should instead incorporate humans as members of the team itself (section 2). Second, even within such teams, we suggest that creative co-creation benefits from configurations in which the user can act as an orchestrator, setting direction, allocating roles among agents, and arbitrating among them over time (section 3). Building on this, we sketch a three-layer agenda spanning the team, interaction, and agent layers (section 4). Finally, we outline open research directions for the AI/ML communities (section 5).

2. From Multi-Agent System (MAS) to Human–Multi-Agent Team (HMAT)

Multi-agent systems (MAS) consist of autonomous agents that solve complex tasks by dividing them into smaller subtasks, each assigned to a specialized agent (Rezaee & Abdollahi, 2015; Dorri et al., 2018; Vig & Adams, 2006). Recent advances in large language models have expanded these possibilities, enabling agents to interpret context, generate novel combinations, and coordinate through natural language rather than predefined protocols (Guo et al., 2024; Wang et al., 2024). Such capabilities allow agents to simulate group dynamics through negotiation, debate, and perspective synthesis.

These properties have made MAS particularly promising for creative work, where problems and solutions often coevolve and decisions simultaneously affect multiple dimensions, requiring diverse, specialized perspectives (Lin et al., 2025; He et al., 2025; Wan & Kalman, 2025). Recent systems illustrate this approach across creative domains. DesignGPT (Ding et al., 2023) assigns product manager and CMF designer agents to different design dimensions, and MARE (Jin et al., 2024) pairs stakeholder, modeler, and checker agents for parallel dependency management.

While MAS shows potential to address the interdependent constraints and iterative cycles of creative work, current implementations have largely overlooked the critical role of sustained human participation. Many studies center on fully autonomous pipelines where humans supply an initial prompt and receive a final output, limiting human involvement to the bookends of the process rather than integrating it throughout the co-evolutionary cycle (Dahiya et al., 2023; Lin et al., 2025). This pattern proves fragile in creative settings, where agents must navigate implicit values, contextual norms, and shifting constraints. Without sustained human guidance, agents may drift from creative intent or optimize toward misaligned objectives (Wang et al., 2024; Li et al., 2023). Realizing the potential of MAS for creative work, therefore, requires active human participation to guide and coordinate agents throughout the process, not merely at its boundaries.

We argue that addressing this gap requires a fundamental reframing. Rather than treating AI agents as tools that humans direct from outside, MAS for creative work should be designed so that humans and agents together constitute a team. The HCI community has long studied this kind of arrangement under the concept of Human-Agent Teams (HATs), in which humans engage as active team members with autonomous AI agents rather than as users of tools (Bansal et al., 2019). Unlike the broader notion of human-AI collaboration, where humans issue commands and agents execute, HATs involve continuous negotiation, role flexibility, and real-time coordination, with humans and agents monitoring each other's states and jointly determining action paths (Jung et al., 2013; Kim et al., 2020; Zhang et al., 2021; 2023; Duan et al., 2024).

Applying this paradigm to MAS yields what we call Human–Multi-Agent Teams (HMATs), defined as teams with at least one human and two or more autonomous agents pursuing shared goals (Lim et al., 2026; Yu et al., 2025; Song et al., 2025; McNeese et al., 2018). Compared with the dyadic HATs that have dominated prior work (Duan et al., 2024; 2025; Walliser et al., 2019; Zhang et al., 2023), HMATs introduce additional interdependencies, since team members must coordinate tasks and integrate outputs across multiple channels (Duan et al., 2025). In particular, inter-agent interactions complicate team design by increasing the risk of unexpected behavior and demanding careful coordination and clear functional boundaries among agents (Naik et al., 2025). While HMATs have begun to attract attention (Dahiya et al., 2023; Iftikhar et al., 2024; Lin et al., 2025), the question of how such teams should be designed for creative work remains largely open. Establishing HMAT as the right unit of analysis does not, by itself, answer this question. Not all HMAT configurations are equally suited to creative work, and identifying which configuration best amplifies human creativity is the question we take up next.

3. HMAT Needs User Orchestration

The composition and structure of a team are widely recognized as central determinants of whether collaborative work succeeds, particularly in creative settings where outcomes depend on how diverse perspectives are integrated (Lappas et al., 2009; Wi et al., 2009; Toh & Miller, 2016). Within human-AI co-creation, prior work has begun to articulate how such interaction should be designed, most notably through frameworks such as COFI (Rezwana & Maher, 2023), which maps the space of interaction patterns between a human and an AI partner. However, this line of work has largely focused on dyadic, one-human-and-one-agent settings, while the multi-agent condition that defines HMATs remains comparatively underexplored. Understanding how HMATs should be configured for co-creation therefore requires empirical grounding in how such teams actually unfold in practice. Studies that give users genuine control over HMAT formation, rather than fixing the team configuration in advance or attending only to dyadic settings, remain scarce. We therefore draw on our previous study in which design practitioners iteratively configured and operated their own HMATs across multiple ideation rounds (Lim et al., 2026), and treat its findings as an illustrative case study rather than as definitive evidence about HMAT design in general.

3.1. Case Study: Understanding Human–Multi-Agent Team Formation for Creative Work

Our prior work (Lim et al., 2026) examined which HMAT formations design practitioners gravitate toward and how those formations shape collaborative design work in practice. To do so, we introduced *CrafTeam*, a technology probe that allows users to actively configure five dimensions of team formation—*team size, structure, role allocation, member composition, and shared mental models*. In a study of 12 design practitioners working in design teams at IT companies, each participant iterated through a cycle of team formation, team ideation, and reflection on the collaboration across three rounds.

According to the study, participants entered expecting to delegate substantial creative work to autonomous agents with minimal human involvement. Most initially designed teams had AI agents collectively perform both productive and reflective roles, with the user delegating the work rather than taking part in it, and some explored configurations in which an AI agent was placed at the top of a hierarchy to lead the others. Across iterations, however, these autonomous configurations produced breakdowns that participants traced to a recurrent limitation. AI agents could readily generate ideas and provide richly worded evaluations, but they did not exercise the value judgments necessary for creative progress. AI evaluators tended to rate nearly all ideas at similar, uniformly high levels, while human evaluators discriminated more sharply between them. AI leaders, when assigned, failed to assign tasks, set a clear direction, or commit to creative choices, and inter-agent interactions tended to fall into unproductive loops, circling around similar concepts rather than making progress.

In response, participants reorganized their teams across cycles. They progressively converged on a configuration in which they themselves served as the team leader and directly coordinated the agents, with a majority of teams across the study adopting a single-tier hierarchy of this form. Participants took on roles that agents could not reliably perform, particularly the reflective role of evaluating ideas, deciding which to develop, and setting the direction for ideation. They also specialized agents into narrower roles, such as idea generation alone or feedback alone, rather than generalist roles, and diversified agent personas to broaden the scope of contributions. As the study reports, users participated in idea evaluation, feedback, and request roles consistently, while taking on idea generation themselves in only a minority of teams, a participation pattern of a coordinator and curator rather than that of a generator.

While this shift broke the unproductive loops, the study reports that it also surfaced new challenges that the system was not designed to address. Participants reported substantial cognitive load from coordinating multiple agents in parallel, since feedback and requests had to be tracked one at a time. They voiced needs for team-level communication channels through which they could direct multiple agents at once, for mechanisms to monitor inter-agent interactions without becoming overwhelmed (and questioned whether agent-to-agent communication should mimic human dialogue at all), and for approaches to team evolution distinct from those used in human teams, since AI agents could be replaced wholesale rather than gradually developed. Together, these surfaced demands suggest that user-orchestrated HMATs are not simply existing co-creation systems with the user nominally in charge. They require deliberate design at multiple layers to be sustainable in practice.

3.2. Why User-Orchestrated HMATs Matter for Co-Creation

This study surfaces user-led orchestration as a recurring organizational pattern in co-creative HMATs. We found that users felt a need to be actively involved in the creative work, and they stayed engaged throughout. Their involvement, however, did not entail producing ideas themselves, and active engagement was separate from idea generation. Instead, they steered the team, setting the direction for ideation, deciding which of the agents' ideas to pursue, and redirecting the agents whenever their exchanges began to circle. This suggests a division of roles between humans and agents in creative HMATs. Agents take on productive and reflective roles, rapidly generating ideas and evaluating them from multiple perspectives, while the human coordinates their work at a meta level.

These findings identify user-orchestrated HMATs as an emerging configuration that warrants further study, a team in which the human user, rather than an autonomous controller, sets direction, coordinates the agents' contributions, and arbitrates among them as the work unfolds. User-orchestration does not require the user to do everything. Agents may still generate ideas, critique proposals, or take initiative within their assigned roles. What distinguishes user-orchestration is that the user keeps the team aligned with the work's purpose as that purpose itself evolves. We offer this not as a settled answer but as a direction worth exploring, one our study points toward without showing that it reliably produces better creative work.

Recent work increasingly argues for keeping humans in an orchestrating role over multi-agent systems, rather than ceding direction to autonomous controllers. Some of this work holds that collaboration is shifting from sequential delegation, where users set goals and then review outputs, toward concurrent co-creation in which the user stays in the loop and keeps directing the work rather than handing it off (Son et al., 2026). Concrete systems built around this idea have also begun to appear. Tools like AgentCoord let users design and revise agent coordination strategies rather than accept an automatically generated one (Pan et al., 2025), while interactive steering tools such as AGDebugger let users edit, reset, and redirect agent teams to keep them on track (Epperson et al., 2025). Our own CrafTeam study echoes this, with practitioners taking up the steering role once they found that agents could neither discriminate sharply among ideas nor direct one another without stalling.

There are reasons to think this configuration is particularly suited to co-creation, even if it may not be the appropriate orientation for every domain. What separates creative co-creation from these neighboring settings is that the work is widely characterized as wicked, with problems and solutions developing together, criteria for success often shifting as the work unfolds, and judgments about quality shaped by the user's evolving purpose and taste (Dorst, 2006). These are exactly the kinds of judgments that human participants bring more reliably than autonomous agents, particularly when agents operate without sustained context about the user's intent. This is precisely the gap practitioners of our previous study stepped in to fill.

However, user-orchestrated HMATs create demands of their own, since such teams work well only when they are carefully designed. Schömbs et al. identify orchestration and conflict resolution as central HCI challenges in interactive multi-agentic systems and call for user-centered design of how people direct such teams (Schömbs et al., 2026). In CrafTeam, designing a user-orchestrated team meant attending not only to team composition but also to interactions suited to the user's coordinating role and to agents specialized to support that role. Putting the user in the orchestrator's seat thus raises deeper questions about how teams should be formed, how humans and agents should interact, and what agents must be capable of. We take up each of these in turn.

4. A Three-Layer Agenda for User-Orchestrated HMATs

Building on this case study and the literature reviewed above, we argue that user orchestration emerged repeatedly in our observations of co-creative multi-agent systems and therefore represents an important research direction. Existing frameworks for co-creation system design map rich design spaces but were not built around an orchestrating user. COFI, for instance, characterizes the interaction design space for a single human working with a single co-creative agent (Rezwana & Maher, 2023), while recent work extends the lens to HMAT design more broadly (Dahiya et al., 2023). Neither, however, addresses how such a design space should shift once the human acts as the team's orchestrator rather than as a prompter or evaluator of its output. To clarify what such a shift requires, we organize our proposed agenda along three layers. The team layer concerns how HMATs are composed and structured; the interaction layer concerns how humans and agents communicate and coordinate during creative work; and the agent layer concerns what individual agents must be able to do to support user-orchestration. Progress at each layer constrains what is achievable at the others, and we therefore see them as a coordinated agenda rather than a list of independent topics.

4.1. Team Layer

Building user-orchestrated HMATs begins with team formation. Existing work has mapped dimensions such as team size and composition (Dahiya et al., 2023; Yu et al., 2025), but user-orchestration imposes a constraint that spans all of them. Regardless of its size, structure, or composition, the team must remain within a single human's capacity to direct. This shapes the team's size, composition, and role allocation.

Team size and structure are limited by the amount of coordination the user can manage, as the orchestration load grows with the number of agents. In CrafTeam, participants reduced the number of agents and converged on single-tier teams they directed themselves, rather than the autonomous, agent-led arrangements they began with. Shrinking the team too far, however, risks losing the diversity of perspectives that creative work benefits from. Persona or role diversity among agents can surface perspectives the user would not have reached alone, but it can also fragment the team's voice into inconsistent or competing directions (Wadinambiarachchi et al., 2024; Shin et al., 2025). Since the user must arbitrate among these contributions, greater diversity broadens exploration while adding to what the user must integrate. How to optimize team size, structure, and composition to keep the team manageable for one user without sacrificing this range remains to be worked out.

Roles must also be divided between the user and the agents, a question of how much of the creative work to automate and how much to leave within the human's domain. Which parts of the creative work stay with the user, especially those that turn on human judgment, and which pass to agents, whether roles hold fixed or shift as the work evolves, and whether the user or the system sets them, each place different demands on the user. This calls for understanding the nature of creative collaboration and the tasks it involves, and how to distribute those tasks across the team, which rests on understanding what agents and humans are each capable of. One direction is to let the user design and adjust the whole arrangement, which maximizes control but adds the burden of setup to the creative work itself. Yet user-orchestration asks the user to direct the work, not to hand-tune every parameter of the team. A more promising path is therefore to let parts of team formation be set or adjusted automatically, within limits fixed by the user's orchestration capacity, so the user can shape the team without being consumed by it.

4.2. Interaction Layer

Team formation specifies who is on the team, but whether user-orchestration holds up in practice depends on how that team interacts during the work. Existing frameworks for human-AI co-creation have mapped a rich space of interaction patterns between a user and an AI partner during the creative task itself. User-orchestrated HMATs introduce a distinct concern that this work does not yet address. Beyond interacting with agents on the creative task, the user must also coordinate and manage the team itself, a form of interaction distinct from co-creation and a rising challenge in its own right (Schömbs et al., 2026).

Once the user is positioned as orchestrator, the interaction layer faces new demands, beginning with how the user works with each agent on its own. The modalities through which the user and an agent share creative artifacts must let the user oversee, intervene in, and refine intermediate states, not only finished outputs. Because the user is managing the agent itself, not only its output, the interaction must also make the agent's ongoing activity legible so the user can follow and steer its work. The communication channels between the user and each agent must in turn support fine-grained direction-setting and feedback throughout the work, beyond initial commands and final acceptance. Prior work that asks when a user should hand a task off to an agent and when they should work alongside it suggests a further need for interaction that lets the user shift between the two as the work unfolds, rather than commit to one at the start (Son et al., 2026). What primitives would make these shifts fluid and let the user re-engage with an agent mid-task without disruption remain largely unexplored.

Because the user now oversees several agents at once, the interaction must let them track each one without becoming overwhelmed, through mechanisms such as activity summaries, state indicators, and lightweight ways to redirect a single agent without disrupting the rest. Recent observations of users with multi-agent teams sharpen these demands (Lim et al., 2026). Users found one-on-one exchanges with each agent insufficient and wanted team-level channels to direct multiple agents or convene them at once. They also found inter-agent dialogue rendered in natural language noisy and costly to follow, and questioned whether agent-to-agent communication needs resemble human conversation at all. Together, these point to interaction primitives that move beyond user-and-AI dialogue toward one-to-many team management and expose inter-agent activity at an abstraction the user can act on, rather than as raw transcript. Investigating such primitives, especially through longitudinal study of how users orchestrate teams over time, is an open agenda for HCI and AI/ML alike, the latter calling for agents that can surface this state directly.

4.3. Agent Layer

In a user-orchestrated HMAT, where the agents carry out the productive work and the user coordinates it, the capabilities of those agents matter all the more. Without complementary design at the agent level, this coordination burden can become unmanageable, and the agents' own contributions to the creative process can go underused. Sustaining the interactions described above thus asks agents to be specialized beyond the ability to perform a particular creative task, toward the role they play as members of a human-led team.

Agents in user-orchestrated HMATs should be specialized not only by role within the creative task, but also for being orchestrated by a human leader rather than running autonomously. This calls for examining what an agent must be able to do in a team context, beginning with how *steerable* it is by the user. Steerability covers how fully an agent can apply the user's requirements and feedback to the task, and how well it can infer the user's implicit intent, the latter of which relates to prior work on theory of mind. It is also supported by agents that surface their reasoning for the user to inspect, that explicitly seek direction at decision points, and that defer when a proposed action falls outside the user's stated intent.

Agents also need forms of *memory* that let them carry context forward throughout the work, so the user is not forced to reissue it or repeatedly correct the same misalignment. Empirical observations of multi-agent teams sharpen this requirement from the user's perspective (Lim et al., 2026).

Practitioners with prior team experience tended to expect AI teammates to grow in capability over time, as human members do, but found this kind of growth ineffective with current agents and instead resorted to wholesale replacement of underperforming ones. Prior work on generative agents has shown how memory streams, reflection, and planning can support behavioral continuity over time (Park et al., 2023). In user-orchestrated HMATs, however, what must persist is not only such continuity but the user's stated direction, the roles they assigned, and the orchestration decisions made along the way. What matters, then, is not how much an agent remembers but how well it captures and retains the user's intent. Beyond the individual agent, the team may also need shared memory through which agents hold common information about the user, connecting to prior work on shared mental models.

Taken together, these observations sharpen the meaning of designing agents for HMATs. The value of an HMAT to the user depends less on raw single-turn capability than on something current agent regimes are not yet centered on, namely an agent's capacity to be directed, to retain that direction, and to remain consistent with a particular user's evolving intent over the course of a creative collaboration. We treat this as a capability requirement at the agent layer, and turn in Section 5.1 to what its pursuit implies for AI/ML research. For ML, this shifts evaluation from single-agent response quality toward role-consistent, user-steerable behavior across multi-agent sessions.

5. Implications and Future Directions

Co-creation in an HMAT shifts the human's role from producing directly to orchestrating multiple agents, a role that cannot pass to the agents themselves, since the judgments creative work turns on, about purpose, value, and direction, are not adequately substituted by autonomous coordination among them. Recognizing this, we have argued that creative HMATs should be user-orchestrated, with the human user remaining the one who keeps the team aligned with the evolving purposes of the work. Realizing such teams requires designing the team, interaction, and agent layers together. HCI research has begun to explore how such teams form and how users should interact with them, while the agents this calls for, built specifically to be orchestrated by a human, fall largely to AI/ML research. We take up this agent layer, setting out concrete directions for the AI/ML research it requires, including what a benchmark for user-orchestrated HMATs would need to capture. We further argue that sustained exchange between the HCI and AI/ML communities will be necessary to advance this agenda.

5.1. Future Directions for AI/ML Research

Meeting the requirements set out at the agent layer would shift the emphasis of AI/ML research. Where much existing work has concentrated on building multi-agent systems and measuring their performance under autonomous task framings, user-orchestrated HMATs call for agents designed expressly for close, sustained interaction with a human in the loop, and for evaluation protocols that can distinguish such agents from their autonomously framed counterparts.

What matters here is not only an agent's task performance but its steerability, how well its behavior aligns with the user's direction over sustained interaction. Prior research has proposed a range of such measures, from how often a user must correct an agent (Chang et al., 2026; Yan et al., 2025; Kosinski, 2024) to whether it holds to stated constraints rather than drifting to defaults (Kwan et al., 2024). These measures, however, were designed for a single agent at a time, and the multi-agent benchmarks that might extend them, such as MultiAgentBench (Zhu et al., 2025), still treat the human as an external task-setter rather than as an orchestrator within the team, scoring the team's output instead of how readily a user can steer it. Closing that gap means first separating an individual agent's steerability from the team's, then measuring the latter in its own right. That calls for benchmarks built around an orchestrating user, ones that fix the user's stated direction and constraints across a sustained, multi-turn exchange and track both how faithfully the team preserves them and how little intervention this demands of the user. Yet even agents that are individually steerable leave the user coordinating the whole team, which points to a different kind of agent worth exploring, a manager agent that sits between the user and the rest of the team and takes on part of the orchestration load.

Recent work has begun to formalize this idea as an autonomous manager that decomposes goals, allocates tasks across human and AI workers, and adapts the workflow on the user's behalf (Masters et al., 2025). For user-orchestrated co-creation, however, the more pressing direction is a manager agent that supports the orchestrating user rather than substituting for them. Rather than contributing to the creative work itself, it would absorb the mechanical coordination of several parallel agents and report back, leaving the user the directional judgments about purpose, value, and which contributions to pursue. Early systems that consolidate low-level agent activity into high-level summaries for human supervision point this way (Zhou, 2025), but the support-oriented manager agent remains underexplored. What needs careful consideration is not the manager's coordination ability but where it draws the line between what it resolves on its own and what it escalates to the user. That line cannot be fixed in advance, since which decisions are mechanical and which are directional shift as the work unfolds. A manager that leans too far toward autonomy reclaims the judgments meant to stay with the user, while one that escalates everything offloads nothing. The boundary itself is what to specify and measure, through how much load it lifts from the user, how rarely it oversteps into the user's judgments, and whether their direction survives the relay to the agents below.

Keeping a team aligned with the user's direction across sessions, not only within one, calls for long-term memory architectures that let agents act as continuing team members rather than stateless tools (Packer et al., 2024). Such memory cannot be a single per-agent buffer. What an agent must keep is not the whole of an interaction but the parts that carry the user's intent and direction. The design target is therefore what an agent retains, not how much it can store. A benchmark would test, after a sustained exchange, whether an agent has kept the right things and can act on them later without the user restating them. Beyond a single agent's memory, the team itself raises a distinct problem. The user's stated direction and the orchestration decisions made so far must be held in common, while task detail can stay local to each agent, a familiar distinction in studies of group memory (Wegner, 1987). Recent work frames multi-agent memory as shared versus distributed stores (Yu et al., 2026), but does not center the orchestrating user. The call, then, is for multi-agent memory designed for user-orchestration, a shared store that captures the user's direction and orchestration decisions and stays under their control. Its success would be measured by whether a team holds to that direction across sessions without re-instruction.

These directions also change how agents should be trained, not only how they are judged. Standard instruction-tuning and RLHF reward a good single response to a prompt (Kwan et al., 2024), which casts the user as a prompter rather than a participant, the very framing user-orchestration sets aside. An orchestrating user instead sets a direction and keeps revising it as the work unfolds. Success then shows up only across a whole session, in whether a team stays on that direction and how seldom the user has to step in to correct it. A training signal for this setting therefore has to span a full multi-turn, multi-agent session and be measured against the user's evolving direction, not a single response to a fixed prompt. That calls for datasets that barely exist today, recorded sessions in which the user's direction and the moments they had to intervene are annotated. We see building these signals and the datasets they need as one of the most concrete near-term contributions ML research can make to user-orchestrated HMATs.

5.2. Integrating HCI and AI/ML Perspectives

Translating this orientation into concrete systems, however, is a problem that no single research community is currently positioned to fully address. HCI research on human-AI co-creation has produced rich accounts of how interaction with AI partners unfolds, but has focused predominantly on dyadic, one-human-and-one-agent settings (Rezwana & Maher, 2023). AI/ML research on multi-agent systems, conversely, has largely been organized around task completion rather than around supporting an orchestrating human (Lin et al., 2025; Dahiya et al., 2023). The space we have outlined in this paper, where humans coordinate multiple agents within a creative team, sits between these efforts and depends on insights from both.

The three-layer agenda is meant to make this collaboration concrete rather than aspirational. The pattern is that what HCI articulates at the team and interaction layers becomes a functional requirement for what AI/ML must build at the agent layer. The team-level channels users asked for in our study, and their difficulty following inter-agent dialogue as raw transcript, become a requirement on the agents themselves, that they surface their activity in a form the user can act on and take direction aimed at the team rather than one channel at a time. AI/ML work runs in the other direction as well, fixing what is feasible at the agent layer and so bounding which team and interaction designs are realistic to pursue. We see progress on user-orchestrated HMATs coming most directly from research that moves between these directions rather than within one.

6. Conclusion

Multi-agent systems have become an increasingly important frame for creative co-creation, but their dominant autonomous orientation risks displacing the human judgments on which creative work most depends. In this paper, we have argued instead for approaching creative co-creation through user-orchestrated Human–Multi-Agent Teams, in which humans participate alongside multiple agents as team members and remain the ones who set direction, coordinate agent contributions, and arbitrate creative decisions as the work's purpose itself evolves. In these teams, the user's role shifts from directly producing the work to making higher-level creative decisions and coordinating the team. Building on this, we developed a three-layer research agenda spanning team formation, interaction design, and agent capabilities. Pursuing this agenda will require sustained collaboration between HCI and AI/ML research, with HCI articulating what user-orchestration concretely requires of teams and interactions, and AI/ML delivering agents that can be directed, can retain that direction, and can remain aligned with a particular user's evolving creative intent over time. We offer this agenda as a starting point for multi-agent co-creation systems that leverage AI agents in service of human creative agency rather than at its expense.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) under the Leading Generative AI Human Resources Development(IITP-2026-RS-2026-25546560) grant funded by the Korea government(MSIT).

References

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. Beyond accuracy: The role of mental models in human-ai team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):2–11, Oct. 2019. doi: 10.1609/hcomp.v7i1.5285. URL https://ojs.aaai.org/index.php/HCOMP/article/view/5285.

Chang, T., Schnabel, T., Swaminathan, A., and Wiens, J. A course correction in steerability evaluation: Revealing miscalibration and side effects in llms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(44):37259–37267, Mar. 2026. doi: 10.1609/aaai.v40i44.41057. URL https://ojs.aaai.org/index.php/AAAI/article/view/41057.

Colton, S., Wiggins, G. A., et al. Computational creativity: The final frontier? In *Ecai*, volume 12, pp. 21–26. Montpelier, 2012.

Dahiya, A., Aroyo, A. M., Dautenhahn, K., and Smith, S. L. A survey of multi-agent human–robot interaction systems. *Robotics and Autonomous Systems*, 161:104335, 2023. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.2022.104335. URL https://www.sciencedirect.com/science/article/pii/S092188902200224X.

Ding, S., Chen, X., Fang, Y., Liu, W., Qiu, Y., and Chai, C. Designgpt: Multi-agent collaboration in design. In *2023 16th International Symposium on Computational Intelligence and Design (ISCID)*, pp. 204–208, 2023. doi: 10.1109/ISCID59865.2023.00056.

Dorri, A., Kanhere, S. S., and Jurdak, R. Multi-agent systems: A survey. *IEEE Access*, 6:28573–28593, 2018. doi: 10.1109/ACCESS.2018.2831228.

Dorst, K. Design problems and design paradoxes. *Design Issues*, 22(3):4–17, 2006. doi: 10.1162/desi.2006.22.3.4.

Duan, W., Zhou, S., Scalia, M. J., Yin, X., Weng, N., Zhang, R., Freeman, G., McNeese, N., Gorman, J., and Tolston, M. Understanding the evolvement of trust over time within human-ai teams. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2), November 2024. doi: 10.1145/3687060. URL https://doi.org/10.1145/3687060.

Duan, W., Flathmann, C., McNeese, N., Scalia, M. J., Zhang, R., Gorman, J., Freeman, G., Zhou, S., Hauptman, A. I., and Yin, X. Trusting autonomous teammates in human-ai teams - a literature review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713527. URL https://doi.org/10.1145/3706598.3713527.

E, J. L., Yen, Y.-C. G., Pan, I. Y., Lin, G., Li, M., Jin, H., Chen, M., Xia, H., and Dow, S. P. When to give feedback: Exploring tradeoffs in the timing of design feedback. In *Proceedings of the 16th Conference on Creativity & Cognition*, C&C '24, pp. 292–310, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704857. doi: 10.1145/3635636.3656183. URL https://doi.org/10.1145/3635636.3656183.

Epperson, W., Bansal, G., Dibia, V. C., Fourney, A., Gerrits, J., Zhu, E. E., and Amershi, S. Interactive debugging and steering of multi-agent ai systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713581. URL https://doi.org/10.1145/3706598.3713581.

Frich, J., Nouwens, M., Halskov, K., and Dalsgaard, P. How digital tools impact convergent and divergent thinking in design ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445062. URL https://doi.org/10.1145/3411764.3445062.

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. Large language model based multi-agents: A survey of progress and challenges. In Larson, K. (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 8048–8057. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/890. URL https://doi.org/10.24963/ijcai.2024/890. Survey Track.

He, J., Treude, C., and Lo, D. Llm-based multi-agent systems for software engineering: Literature review, vision, and the road ahead. *ACM Trans. Softw. Eng. Methodol.*, 34(5), May 2025. ISSN 1049-331X. doi: 10.1145/3712003. URL https://doi.org/10.1145/3712003.

Iftikhar, R., Chiu, Y.-T., Khan, M. S., and Caudwell, C. Human–agent team dynamics: A review and future research opportunities. *IEEE Transactions on Engineering Management*, 71:10139–10154, 2024. doi: 10.1109/TEM.2023.3331369.

Jin, D., Jin, Z., Chen, X., and Wang, C. Mare: Multi-agents collaboration framework for requirements engineering, 2024. URL https://arxiv.org/abs/2405.03256.

Jung, M. F., Lee, J. J., DePalma, N., Adalgeirsson, S. O., Hinds, P. J., and Breazeal, C. Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pp. 1555–1566, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450313315. doi: 10.1145/2441776.2441954. URL https://doi.org/10.1145/2441776.2441954.

Khan, A., Shokrizadeh, A., and Cheng, J. Beyond automation: How designers perceive ai as a creative partner in the divergent thinking stages of ui/ux design. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713500. URL https://doi.org/10.1145/3706598.3713500.

Kim, S., Eun, J., Oh, C., Suh, B., and Lee, J. Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pp. 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376785. URL https://doi.org/10.1145/3313831.3376785.

Kosinski, M. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024. doi: 10.1073/pnas.2405460121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2405460121.

Kwan, W.-C., Zeng, X., Jiang, Y., Wang, Y., Li, L., Shang, L., Jiang, X., Liu, Q., and Wong, K.-F. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20153–20177, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1124. URL https://aclanthology.org/2024.emnlp-main.1124/.

Lappas, T., Liu, K., and Terzi, E. Finding a team of experts in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pp. 467–476, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584959. doi: 10.1145/1557019.1557074. URL https://doi.org/10.1145/1557019.1557074.

Li, G., Hammoud, H., Itani, H., Khizbullin, D., and Ghanem, B. Camel: Communicative agents for "mind" exploration of large language model society. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008. Curran Associates, Inc., 2023.

Lim, H., Choi, D., Nam, S., Kim, B., and Hong, H. Understanding human–multi-agent team formation for creative work. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3791166. URL https://doi.org/10.1145/3772318.3791166.

Lin, Y.-C., Chen, K.-C., Li, Z.-Y., Wu, T.-H., Wu, T.-H., Chen, K.-Y., Lee, H.-y., and Chen, Y.-N. Creativity in LLM-based multi-agent systems: A survey. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27584–27607, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1403. URL https://aclanthology.org/2025.emnlp-main.1403/.

Masters, C., Vellanki, A., Shangguan, J., Kultys, B., Gilmore, J., Moore, A., and Albrecht, S. Orchestrating human-ai teams: The manager agent as a unifying research challenge. In *Proceedings of the 2025 7th International Conference on Distributed Artificial Intelligence*, DAI '25, pp. 91–107, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400722752. doi: 10.1145/3772429.3772439. URL https://doi.org/10.1145/3772429.3772439.

McNeese, N. J., Demir, M., Cooke, N. J., and Myers, C. Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human Factors*, 60(2):262–273, 2018. doi: 10.1177/0018720817743223. URL https://doi.org/10.1177/0018720817743223. PMID: 29185818.

Naik, S., Toombs, A. L., Snellinger, Amanda, P., Saponas, S., and Hall, A. K. Designing with multi-agent generative ai: Insights from industry early adopters. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, DIS '25, pp. 1961–1972, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714856. doi: 10.1145/3715336.3735823. URL https://doi.org/10.1145/3715336.3735823.

Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S. G., Stoica, I., and Gonzalez, J. E. Memgpt: Towards llms as operating systems, 2024. URL https://arxiv.org/abs/2310.08560.

Pan, B., Lu, J., Wang, K., Zheng, L., Wen, Z., Feng, Y., Zhu, M., and Chen, W. Agentcoord: Visually exploring coordination strategy for llm-based multi-agent collaboration. *Computers & Graphics*, 132:104338, 2025. ISSN 0097-8493. doi: https://doi.org/10.1016/j.cag.2025.104338. URL https://www.sciencedirect.com/science/article/pii/S0097849325001797.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Rezaee, H. and Abdollahi, F. Average consensus over high-order multiagent systems. *IEEE Transactions on Automatic Control*, 60(11):3047–3052, 2015. doi: 10.1109/TAC.2015.2408576.

Rezwana, J. and Maher, M. L. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Trans. Comput.-Hum. Interact.*, 30(5), September 2023. ISSN 1073-0516. doi: 10.1145/3519026. URL https://doi.org/10.1145/3519026.

Schömbs, S., Zhang, Y., Goncalves, J., and Johal, W. From conversation to orchestration: Hci challenges and opportunities in interactive multi-agentic systems. In *Proceedings of the 13th International Conference on Human-Agent Interaction*, HAI '25, pp. 158–168, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400721786. doi: 10.1145/3765766.3765795. URL https://doi.org/10.1145/3765766.3765795.

Shin, J., Polyanskaya, A., Lucero, A., and Oulasvirta, A. No evidence for llms being useful in problem reframing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713273. URL https://doi.org/10.1145/3706598.3713273.

Shneiderman, B. Creativity support tools. *Communications of the ACM*, 45(10):116–120, 2002.

Son, K., Lee, H., Choi, D., Kim, Y., Kim, T. S., Lee, Y., Chung, J. J. Y., Jung, H., and Kim, J. "when to hand off, when to work together": Expanding human-agent co-creative collaboration through concurrent interaction, 2026. URL https://arxiv.org/abs/2603.02050.

Song, T., Tan, Y., Zhu, Z., Song, M., Yibin, F., and Lee, Y.-C. The more, the stronger? investigating how multi-agent AI shapes human opinions. In *ICLR 2025 Workshop on Human-AI Coevolution*, 2025. URL https://openreview.net/forum?id=6zlttMWe4G.

Toh, C. A. and Miller, S. R. Creativity in design teams: the influence of personality traits and risk attitudes on creative concept selection. *Research in Engineering Design*, 27(1):73–89, Jan 2016. ISSN 1435-6066. doi: 10.1007/s00163-015-0207-y. URL https://doi.org/10.1007/s00163-015-0207-y.

Vig, L. and Adams, J. Multi-robot coalition formation. *IEEE Transactions on Robotics*, 22(4):637–649, 2006. doi: 10.1109/TRO.2006.878948.

Wadinambiarachchi, S., Kelly, R. M., Pareek, S., Zhou, Q., and Velloso, E. The effects of generative ai on design fixation and divergent thinking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642919. URL https://doi.org/10.1145/3613904.3642919.

Walliser, J. C., de Visser, E. J., Wiese, E., and Shaw, T. H. Team structure and team building improve human–machine teaming with autonomous agents. *Journal of Cognitive Engineering and Decision Making*, 13(4):258–278, 2019. doi: 10.1177/1555343419867563. URL https://doi.org/10.1177/1555343419867563.

Wan, Y. and Kalman, Y. M. Using generative ai personas increases collective diversity in human ideation, 2025. URL https://arxiv.org/abs/2504.13868.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., and Wen, J. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18

(6):186345, Mar 2024. ISSN 2095-2236. doi: 10.1007/s11704-024-40231-1. URL https://doi.org/10.1007/s11704-024-40231-1.

Wegner, D. M. *Transactive Memory: A Contemporary Analysis of the Group Mind*, pp. 185–208. Springer New York, New York, NY, 1987. ISBN 978-1-4612-4634-3. doi: 10.1007/978-1-4612-4634-3_9. URL https://doi.org/10.1007/978-1-4612-4634-3_9.

Wi, H., Oh, S., Mun, J., and Jung, M. A team formation model based on knowledge and collaboration. *Expert Systems with Applications*, 36(5):9121–9134, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2008.12.031. URL https://www.sciencedirect.com/science/article/pii/S095741740800907X.

Xu, X. T., Konnova, A., Gao, B., Peng, C., Vo, D., and Dow, S. P. Productive vs. reflective: How different ways of integrating ai into design workflows affect cognition and motivation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713649. URL https://doi.org/10.1145/3706598.3713649.

Yan, J., Luo, Y., and Zhang, Y. Refutebench 2.0 – agentic benchmark for dynamic evaluation of llm responses to refutation instruction. *IEEE Transactions on Audio, Speech and Language Processing*, 33:4760–4770, 2025. doi: 10.1109/TASLPRO.2025.3626926.

Yu, Z., Cui, T., Luo, C., and Tan, D. A systematic literature review on human-agent teaming with insights into multi-agent interactions. In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS) 2025*, volume 20, 2025. URL https://aisel.aisnet.org/pacis2025/hci/hci/20. Track 5: Human Computer Interaction, Paper Number: PACIS2025-1614.

Yu, Z., Yu, N., Zhang, H., Ni, W., Yin, M., Yang, J., Zhao, Y., and Zhao, J. Multi-agent memory from a computer architecture perspective: Visions and challenges ahead, 2026. URL https://arxiv.org/abs/2603.10062.

Zhang, R., McNeese, N. J., Freeman, G., and Musick, G. "an ideal human": Expectations of ai teammates in human-ai teaming. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3), January 2021. doi: 10.1145/3432945. URL https://doi.org/10.1145/3432945.

Zhang, R., Duan, W., Flathmann, C., McNeese, N., Freeman, G., and Williams, A. Investigating ai teammate communication strategies and their impact in human-ai teams for effective teamwork. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), October 2023. doi: 10.1145/3610072. URL https://doi.org/10.1145/3610072.

Zhou, J. Orchvis: Hierarchical multi-agent orchestration for human oversight, 2025. URL https://arxiv.org/abs/2510.24937.

Zhu, K., Du, H., Hong, Z., Yang, X., Guo, S., Wang, Z., Wang, Z., Qian, C., Tang, X., Ji, H., and You, J. MultiAgentBench : Evaluating the collaboration and competition of LLM agents. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8580–8622, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.421. URL https://aclanthology.org/2025.acl-long.421/.