Developing Guidelines for Human-LLM Agent Teams: A Multi-Stakeholder Lens

Mireia Yurrita
Utrecht University
Utrecht, The Netherlands
m.yurritasemperena@uu.nl

Davide Dell'Anna
Utrecht University
Utrecht, The Netherlands
d.dellanna@uu.nl

Pradeep K. Murukannaiah
Delft University of Technology
Delft, The Netherlands
p.k.murukannaiah@tudelft.nl

Catholijn M. Jonker
Delft University of Technology
Delft, The Netherlands
c.m.jonker@tudelft.nl

Pınar Yolum
Utrecht University
Utrecht, The Netherlands
p.yolum@uu.nl

ABSTRACT

Agents based on Large Language Models (LLM agents) have the potential to work with humans as part of a team to achieve specific goals. The natural language interface of LLM agents and their high level of autonomy enables more seamless collaborations than previous technologies, allowing them to carry out tasks autonomously and engage in conversations with humans, e.g., to clarify goals, request authorizations, or double-check decisions. However, the current literature lacks systematic design guidelines for these human-LLM agent teams. This gap might foster misunderstandings, misuse of autonomy, and lack of common ground, potentially leading to collaboration pitfalls. To mitigate these risks, we develop 24 guidelines for the principled design of human-LLM agent teams. We adopt a multi-stakeholder approach and propose guidelines for LLM agents, human team members, team designers and embedding organizations. To develop these guidelines, we distill design recommendations from an exploratory workshop with 15 experts on human-AI teaming and a literature review of 93 empirical papers in human-LLM collaboration. Drawing from literature on human teams, we conceptually categorize the recommendations across different stages of the teaming process. A user study with 10 additional experts suggests the guidelines can help prevent collaboration pitfalls in human-LLM agent teams within workplace settings.

KEYWORDS

human-LLM agent teams; AI agents; generative AI; hybrid intelligence; guidelines

ACM Reference Format:
Mireia Yurrita, Davide Dell'Anna, Pradeep K. Murukannaiah, Catholijn M. Jonker, and Pınar Yolum. 2026. Developing Guidelines for Human-LLM Agent Teams: A Multi-Stakeholder Lens . In *Proc.of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 11 pages. https://doi.org/10.65109/JOWO4591

Proc.of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/JOWO4591

1 INTRODUCTION

Human-Artificial Intelligence (AI) teams are collaborative units where humans and AI systems cooperate towards a common goal [46]. For AI systems to be considered team members, they need to exhibit partial to high levels of autonomy and be able to coordinate joint activity [72]. As AI systems get equipped with the ability to perceive and interact with users, autonomously act on their surroundings, and continuously learn and adapt [21], they are becoming capable of synergistically working as part of human-AI teams [26, 85]. Successful human-AI teams benefit from the efficiency of AI systems and the flexibility of humans [1, 30, 101].

Recent AI developments have taken the form of agents powered by Large Language Models (LLM). LLM agents are capable of executing tasks on behalf of humans [71], e.g., scheduling meetings or ordering groceries [10]. Two characteristics make LLM agents distinct from previous technologies and particularly suitable for human-AI teaming. First, LLM agents can engage in dialogues through natural language with their human counterparts [10, 71]. Natural language communication provides a shared linguistic foundation, facilitating the establishment of common ground between humans and LLM agents [49, 79]. Second, LLM agents can plan and carry out workflows by invoking a range of external tools [71]. This enables LLM agents to coordinate tasks, accept delegated responsibilities and execute complex workflows in(ter)dependently [71].

Although LLM agents show potential for teaming with humans, the design of human-LLM agent teams requires guidance to ensure effective teamwork and to avoid potential collaboration pitfalls [80]. Prior work has suggested systematic methods for the design of trustworthy AI [97], human-AI interactions [4], human-LLM interactions [58, 96], or human-AI teams [16, 90]. However, none of these methods specifically focus on human-LLM agent teams, nor do they address the challenges that the distinct nature of LLM agents might bring to teaming processes.

To address this gap, we develop 24 guidelines for human-LLM agent teaming. Within team settings, guidelines constitute sets of principles aimed at promoting best practices in team development and training [80]. We highlight the temporal dimension of teaming by arranging our guidelines across different stages of the teaming process: early stages, active collaboration stage, completion stage, continuous learning and affect management [43, 61]. Since human-LLM agent teams bring opportunities and challenges across stakeholders [83], we adopt a multi-stakeholder perspective,

offering guidance for LLM agents, human team members, team designers, and the organization where the team is embedded.

We used an iterative process (described in Table 1) to conceive, consolidate and refine our human-LLM agent teaming guidelines. We first ran a workshop with 15 human-AI teaming experts to capture their initial ideas and desiderata for human-LLM agent teaming guidelines. We then conducted a literature review to identify lessons from empirical research. We consolidated insights from the exploratory workshop and literature into 27 guidelines. Finally, we refined those guidelines through a user study with 10 additional experts, resulting in 24 final guidelines.

Our paper, therefore, makes the following contributions:

- We systematically develop a set of 24 guidelines for human-LLM agent teams. These guidelines explicitly account for the temporal dimension of teaming processes and adopt a multi-stakeholder perspective on teaming. Each guideline includes strategies to operationalize the guideline.
- We evaluate the relevance, clarity and generative power (i.e., ability to help stakeholders think of actions for the principled design of teams) of our guidelines. Our findings suggest that the guidelines can help prevent human-LLM agent collaboration pitfalls in workplace settings.

Section 2 summarizes related work. Section 3 provides an overview of the guidelines. Sections 4, 5 and 6 detail the three phases of the iterative development of the guidelines. Section 7 discusses the role the AAMAS community plays in the guidelines realization.

2 RELATED WORK

Human-AI teams. Recent research has used the term *human-AI teams* to describe collaborative ensembles where humans and AI systems work interdependently towards a common goal [72]. Human-AI teams have also been referred to as human-autonomy, human-agent, or hybrid intelligence teams [28, 40, 47, 55]. Human-AI teams include at least one human working cooperatively with at least one autonomous agent that presents partial or high-degree of self-governance [28, 66, 67, 72]. According to O'Neill et al. [72], for agents to be part of a human-AI team they must meet at least partial levels of autonomy, i.e., level 5 or 6 in the Levels of Autonomy (LOA) continuum [74]. This allows agents to "fulfill a distinct role in the team and make a unique contribution to performance" [57].

Research on human-AI teams has explored aspects such as trust and explanations [9, 53, 65, 94, 103], team formation and cohesion [35, 92], proactivity and shared mental models [29, 36], and interdependence in team functioning and team performance [45, 50, 75]. For instance, Van Zoelen et al. [91] present a human-centered approach for the specification of design solutions for human-AI teams and their abstraction into team design patterns. Centeio Jorge et al. [16] introduce the Interdependence and Trust Analysis (ITA) framework to support the design of human-machine teams. Dell'Anna et al. [26] propose attributes of effective human-AI teams, such as communication, coordination, proactivity, and awareness and shared knowledge of team structure, member skills, team objectives, and team norms. These works are especially relevant in the era of LLM agents, yet they do not capture the challenges that increased levels of autonomy and a natural language interface might bring to human-LLM agent teams.

Existing Guidelines. Prior work on human-centered AI has proposed guidelines for the principled design of Trustworthy AI and human-AI interaction (e.g., [4, 58, 96, 97]). For example, Wickramasinghe et al. [97] proposed guidelines for Trustworthy AI development by addressing interactions between (1) AI systems and developers, (2) AI systems and users, and (3) developers and users. Amershi et al. [4] provided 18 guidelines to design AI systems that lead to desirable human-AI interactions. Weisz et al. [96], focused on the interaction between humans and generative AI and suggested 6 design principles and 32 strategies for the design of generative AI applications. Lin et al. [58] suggested 11 guidelines for human-generative AI interaction, each referring to key moments of the user experience.

Despite the progress, none of these guidelines focus on human-LLM agent teams. Such guidelines are, in turn, important to ensure effective teamwork among humans and LLM agents and to avoid collaboration pitfalls [80]. Motivated by this gap, we draw from literature on human teams and adopt a team-oriented perspective on the interplay between humans and LLM agents.

Human teams are defined by the interdependent work of two or more members who share a collective identity and a common goal [81]. Decades of research have generated theoretical models of teams, where teams are presented as "complex, adaptive, dynamic systems" [64]. Marks et al. [61] suggested a conceptual model where team processes unfold in recurring stages during an episode of team performance. They distinguished three types of activities that characterize team processes. *Transition processes* involve developing, adapting, and clarifying the team's common purpose, strategy, tasks, and role structure. *Action processes* involve back-up behaviors, mutual performance monitoring, monitoring of goal progression, and coordination. *Interpersonal processes* involve affect management, conflict management, and motivational issues. Ilgen et al. [43] propose a teaming model that complements Marks et al. [61]'s framework, and in addition to the teams' active collaboration period, also considers the early stages and finishing stages of teaming processes, capturing one episode in the developmental cycle of a team. Ilgen et al. [43]'s model falls within the Input-Mediator-Output (IMO) family of models [44, 63]. In IMO models, *team Inputs* (team composition, tasks complexity, members' differences), are converted by *team Mediators* into *team Outputs* (team viability, individual learning, development and satisfaction [39]). Mediators include team processes [11, 62] and emergent states (i.e., affective or cognitive qualities of the team, such as trust and shared mental models.) [24, 44, 51].

In this paper, we use Marks et al. [61]'s and Ilgen et al. [43]'s models as a theoretical lens to organize our guidelines. We align our guidelines with the temporal dimension presented in these models, yet adapt them to capture new strategies that human-LLM agent teams might necessitate.

3 GUIDELINES OVERVIEW

Figure 1 provides an overview of the 24 guidelines for human-LLM agent teams we developed. While a few guidelines are broad enough to be applicable to other human-agent teams, the presented collection of 24 guidelines account for the distinctive features that

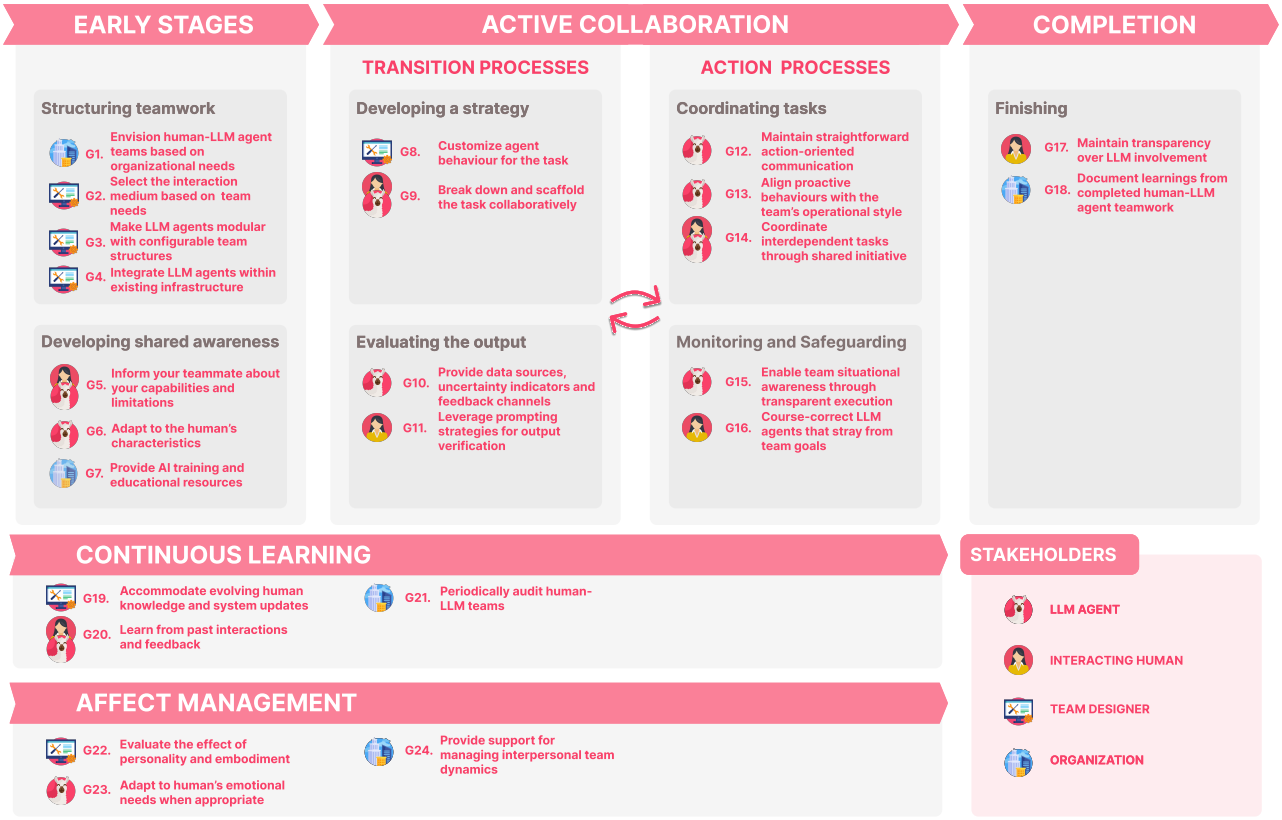

Figure 1: Graphical representation of our 24 guidelines. Guidelines are arranged across five main team development stages: early stages, active collaboration, completion, continuous learning, affect management. Each guideline is directed at one or more stakeholders: the LLM agent, interacting human, team designer or organization.

make human-LLM agent teams especially promising. These guidelines focus on guiding different stakeholders at an organization to take actions that contribute to effective teamwork and that might prevent collaboration pitfalls between humans and LLM agents. Each guideline is written as a rule of action, containing 4-8 words [4] and is followed by specific strategies that help adhere to the guideline—the detailed version of the guidelines and their provenance can be found in the Appendix within the supplementary material: https://osf.io/75z83/.

We organize our guidelines using a process-oriented framework that explicitly accounts for the temporal dimension of teaming [43, 61]. Specifically, the guidelines cover one complete episode in the human-LLM agent team developmental cycle [43] and are structured around five main team development stages. (1) **Early stages** of team development [43]: Guidelines focus on structuring teamwork and building shared awareness between humans and LLM agents. (2) **Active collaboration** stage: As the team develops and starts working together towards a common goal [43], human-LLM agent teams engage in two main process types [61]. Through **transition processes** the team develops a strategy and evaluates the output. Through **action processes** the team coordinates tasks and monitors progress. We visualized transition and action processes as cyclical to indicate the dependencies between them: the output of a transition process can become the input of an action process and vice versa [61]. Transition and action processes can happen simultaneously or sequentially. (3) **Completion stage**: Teams complete one episode in their developmental cycle and prepare to begin a new cycle [43]. Throughout all stages, guidelines for (4) **Continuous learning** [43] and (5) **Affect management** [61] remain active, supporting the team's ongoing development and effectiveness.

We direct the guidelines at four main actors. First, the **LLM agent** itself. Due to increased autonomous capabilities of LLM agents [52] and the ability to influence their behaviors, we treat LLM agents as distinct team members. Guidelines for LLM agents can be e.g. embedded through reinforcement learning from human feedback (RLHF) during training or injected as system prompts during design time, so that the LLM agent behavior follows such guidelines during execution time. For example, if G12 is injected as a system prompt during design time, the LLM agent can maintain an action-oriented communication style during execution time. We treat the foundation model underlying LLM agents as an on-demand component: organizations can fine-tune the foundational model and can build infrastructure around it (e.g., interface, integration with other tools), but are not responsible for the development of the foundation model itself. We assume that the development of the

foundation model (not the LLM agent) happens prior to the early stages of teaming and is, therefore, out of the scope of this paper. Guidelines directed at LLM agents assume that the LLM has the technical capabilities to follow the provided guidelines.

Second, **human(s) interacting** with the LLM agent. Guidelines directed at the interacting human provide principled ways for the human to share context, develop shared mental models and coordinate actions with the LLM agent [37]. Guidelines at later stages of the teaming process depend on the knowledge that the human develops in earlier stages, e.g., if the organization provides training for humans to craft effective natural language prompts (guideline 7, G7) the human is capable of leveraging prompting strategies for output verification (G11).

Third, **team designers**. We use *team designer* to denote the equivalent of a team manager in human teams. Team designers have the knowledge to understand the contextual needs of human-LLM agent teams (G2) and to structure these teams by e.g., making LLM agents modular (G3). In real-world organizations, the role of team designers might be distributed across different actors (e.g., development team, project manager).

Fourth, the **organization** where human-LLM agent teams are embedded. These guidelines evoke actions for higher management to support the teaming process, e.g., periodically auditing human-LLM teams (G21).

4 PHASE 1: EXPLORATORY WORKSHOP

We first report on a workshop that we conducted with 15 experts in human-AI teaming to identify their (1) *initial ideas* and (2) *desiderata* towards human-LLM agent teaming guidelines.

4.1 Workshop Procedure

Participants were organized in three groups. They were first shown a video[1] of Open AI Operator[2], an LLM agent based on ChatGPT that can perform complex web tasks on behalf of users. Operator is a good example of LLM agents that could engage in teamwork with humans, thanks to its partial autonomy, and its abilities to interact through natural language, and to engage in tasks interdependently with the human. Participants were asked to imagine they were working in a company trying to build an LLM agent like Operator but with enhanced capabilities for teaming. They were given an initial set of design guidelines for human-AI interaction [4] (focused on the interaction medium) and human-agent collaboration [20] (focused on interdependence and coordination) and asked to turn these generic guidelines into specific system requirements for the engineering and UX design teams. Through this task, we wanted participants to ideate around characteristics required from humans and LLM agents for effective teaming. We finally provided participants with the quality model by Dell'Anna et al. [26] for hybrid intelligence teams. We asked them to evaluate the extent to which the conceived system requirements would contribute to each of the dimensions in the quality model. Through this task we wanted participants to specifically reflect on whether and how humans and LLM agents can be part of a team and what would be required for guidelines to capture this.

[1] https://www.youtube.com/watch?v=gYqs-wUKZsM (last accessed 07.10.2025)
[2] https://openai.com/index/introducing-operator/ (last accessed 07.10.2025)

Participant Recruitment. For the recruitment of participants, we used convenience sampling. We advertised the workshop through relevant mailing and Slack channels. Out of the 15 participants, 8 identified as experts in AI-related topics, including multi-agent systems, neurosymbolic systems, natural language processing, knowledge representation, or reinforcement learning. The other 7 participants specialized in human-computer interaction, human-agent interaction, human-AI teams and sociotechnical systems. The participation in our workshop was voluntary. Our study was allowed to proceed by the research committee of our institution on the basis of an *Ethics and Privacy Quick Scan*.

Data Collection and Analysis. The data we generated consists of notes created by participants in the workshop activities and transcripts of the workshop discussions. We analyzed the data using *reflexive thematic analysis* [12, 23]. The transcripts were analyzed in the following way: (1) we had the audio recordings automatically transcribed, (2) we read the transcriptions and (3) grouped quotes in codes and code groups. (4) The first author crafted the first set of themes. (5) All authors reviewed the themes and (6) refined the codes. The themes resulting from the workshop include initial ideas and desiderata for human-LLM agent teaming guidelines.

4.2 Results of the Exploratory Workshop

We discuss the main insights generated in the exploratory workshop below. We refer to quotes as W-P*i*. W refers to *workshop* and P to *participant*, *i* being the index of each quoted participant.

Initial ideas. Experts' initial ideas for human-LLM agent teaming guidelines revolved around five main themes.

Transparency. Participants highlighted the role of transparency in enabling LLM agents engage in teaming. This included the need for LLM agents to communicate their capabililities to the user at the beginning of the teaming process, or the need to provide step-by-step explanations during execution. Explanations, however, were not considered always necessary, *"You don't want to [get] an explanation for every click"* (W-P2). Participants also mentioned the need for LLM agents to refer to their sources, and disclose how trustworthy these were. Transparency was also discussed in relation to accessibility, which participants related to having a clean design as well as a user-friendly interface that would minimize the effort from humans to understand what the LLM agent communicates.

Proactivity. Proactivity of LLM agents was considered especially important in cases where the uncertainty level of the output was high. In those cases, LLM agents would need to explicitly consult the user (e.g., *"If it's unclear if something is gluten free or not, that is a moment to check with the user"* (W-P8)). Correct timing of proactive behaviors was considered key. Participants also referred to the need for the LLM agent to be an active listener and to adapt to human preferences.

Coordination and control. While most of the ideas revolved around LLM agent capabilities (the main focus of previously suggested design guidelines [4, 20]), experts also discussed ideas related to interaction and coordination. They highlighted the need for humans and LLM agents to align their understanding of the situation and to coordinate their actions. Humans were considered responsible for setting preferences for the LLM agent actions and

Table 1: Overview of the iterative process for the conception, consolidation and evaluation of guidelines.

#	Phase	Objective	Method	Use case	Data Analysis	Outcome
1	Exploratory workshop	Identify experts' initial ideas and desiderata for guidelines	Workshop with 15 experts in human-AI teaming arranged in three groups	Open AI Operator	Reflexive Thematic Analysis	Initial ideas about transparency, proactivity, coordination, control, safeguards and adaptivity. Three recommendations for the guideline development phase.
2	Guideline development	Conceive and consolidate guidelines	Literature review of 93 papers and combination of literature with workshop insights for guideline consolidation	Agnostic	Reflexive Thematic Analysis	27 guidelines directed at four stakeholders and arranged across five team development stages.
3	Evaluation and refinement	Evaluate and refine guidelines	Survey and think-aloud sessions with 10 additional experts in human-AI teaming	Personal health, coding, travel, legal assistant	Reflexive Thematic Analysis	24 refined guidelines. Ratings of relevance and clarity per guideline group. Experts' reflections on guidelines.

for constraining its abilities through interaction. However, participants acknowledged that, even if the human should be the one who is globally in control, in a mixed initiative interaction, there are moments when *"the control is shifted"* (W-P4). This requires team members to ensure proper control handover at specific moments.

Safeguards. For LLM agents like Operator, participants considered several safeguards to limit potential risks. These included active consent requests or access restrictions to e.g., unsafe websites or the hard drive. Participants also highlighted that LLM agents should not be allowed to proceed with actions that involved making payments or modifying files (e.g., *"A safeguard that you would definitely want is that it doesn't modify your files"* (W-P11)).

Adaptivity. Workshop discussions addressed the temporal dimension of teaming. From the LLM agent's perspective, the need to adapt to user preferences over time was highlighted. Similarly, agent proactivity and safeguards might also evolve over time. Participants highlighted the tendency to think of one-shot interactions and the need to *"talk about a system that learns over time"* (W-P1).

Desiderata. From the workshop, we distilled three main expert recommendations to inform the guideline conception phase.

Desideratum 1: *Guidelines for human-LLM agent teams should explicitly account for team members other than the LLM agent.* When workshop participants contrasted the provided design guidelines [4, 20] with the quality model for hybrid intelligence teams [26], they noticed that the guidelines mostly made an emphasis on AI capabilities. However, they lacked pointers as to e.g., how the human could adapt to the agent's behavior as the teaming process evolves. Experts, therefore, indicated that human-LLM agent teaming guidelines should more concretely involve the human team member and the context where the teaming happens.

Desideratum 2: *Guidelines for human-LLM agent teams should account for the intrinsically asymmetric relation between humans and LLM agents.* For this reason, they suggested that using a unit of analysis at a team level (as in the quality model by Dell'Anna et al. [26]) might not be helpful. Instead, guidelines would benefit from shifting the unit of analysis to the level of team members, enabling organizations to account for the differences in behaviors and requirements between humans and LLM agents.

Desideratum 3: *Guidelines for human-LLM agent teams should point to different temporal dimensions of teaming.* While the temporal dimension is present in human-AI interaction guidelines [4], it is not prevalent in human-agent collaboration considerations [20]. This was considered an important feature to capture the way interdependencies among team members evolve over time.

5 PHASE 2: GUIDELINE DEVELOPMENT

In the second phase, we conducted a systematic literature review to identify lessons learned from empirical research in human-LLM collaboration. Insights from literature were combined with the initial ideas provided by the experts in phase 1 (section 4), and arranged in accordance with expert desiderata. Since empirical research in human-LLM agent teaming is scarce (Operator was released in January 2025), we more broadly searched for lessons learned from human-LLM collaboration and drew special attention to research directions suggested by the authors for future scenarios where LLMs would be provided with higher levels of autonomy. This approach is in line with the study conducted by Cila [20].

5.1 Systematic Literature Review

We conducted the literature search on the ACM Digital Library (https://dl.acm.org/). We chose ACM Digital Library because it represents a comprehensive source of papers published in major venues on multi-agent systems as well as human-computer and human-robot interaction. The choice of limiting the search to ACM Digital Library provided us with a dataset that focuses on the field of computing and avoids specialized topics that are out of the scope of this paper [33]. We began our literature review by identifying appropriate search terms to ensure wide coverage of empirical research on human-LLM collaboration [28, 72], which led to an initial set of 241 papers (as of 17 July 2025). The search terms are listed in the Appendix.

We then conducted a manual screening of the title and abstract of the resulting 241 papers. We defined seven exclusion criteria to only focus on full papers or late-breaking work presenting empirical research where at least one human and one LLM or agent would communicate in natural language—see Appendix. Following the application of exclusion criteria (EC), 93 papers remained for review. We then inspected the shortlisted 93 papers systematically and extracted design recommendations from them. For each paper, we extracted the author, year and venue of publication, title, objective of the paper, studied context, relevant results and design recommendations. In some cases, design recommendations were directly provided by authors. In those cases, we directly included the provided design recommendations in our data extraction template

(e.g., we included "*Visualize a pipeline while the user is prompting the system*" [105] as a design recommendation). In the papers where design recommendations were not directly provided by the authors, we distilled recommendations based on relevant results (e.g., from "*our qualitative findings suggest AI characters should have emotional intelligence*"[99] we derived a design recommendation in the form of: "*Embed emotional intelligence in AI characters*"). Data extraction led to a total of 481 (non-unique) design recommendations for clustering and consolidation.

5.2 Consolidation of Guidelines

For guideline consolidation, we combined the initial ideas generated in the expert workshop and the 481 design recommendations distilled from literature. We engaged in an iterative process of *reflexive thematic analysis* [22, 23] and arranged our guidelines following the expert desiderata distilled from the exploratory workshop. Reflexive thematic analysis was chosen because: (1) it allows the interpretive integration of multiple qualitative data sources (i.e., workshop and literature findings), and (2) the method's iterative and reflexive nature enables us to construct meaning across different data types and translate these insights into practice-oriented guidelines. The first author took the lead in guideline consolidation. All authors contributed to the consolidation of guidelines through the review and further iteration of coding results.

First, we inductively coded all design recommendations and reached data saturation, i.e., there was a point where design recommendations would repeat what was already known through the workshop or from papers reviewed earlier. This first inductive phase aimed at coding and grouping together design recommendations with extensive overlap into higher-level design considerations. The *second* round of coding was deductive in nature. Following expert desiderata 1 and 2, we assigned each design consideration to the corresponding stakeholder based on which stakeholder was better suited at realizing the design consideration: LLM agent, interacting human, team designer or the embedding organization. In the *third* round, individual codes were inductively clustered into themes based on purpose similarity, e.g., codes focused on developing situational awareness. The *fourth* round of coding was deductive. We mapped each theme to the corresponding team development stage [43, 61]. For this mapping, we identified the strongest association between the themes and the definition of each team development stage [43, 61]. The thematic analysis led to an initial set of 27 guidelines arranged in 5 team development stages i.e., early stages, active collaboration stage, completion stage, continuous learning and affect management. This arrangement of guidelines differentiates team design time from execution time, which aligns with expert desideratum 3 from the exploratory workshop. After the thematic analysis, we refined the guidelines through 5 iterations involving all authors. We articulated guidelines as a rule of action, followed by a description and strategies for adhering to such guidelines. The detailed version of the guidelines is shown in the Appendix.

6 PHASE 3: EVALUATION AND REFINEMENT

In the third phase, we conducted a two-step user study to evaluate and refine our guidelines. The objective was to provide an initial assessment of the capacity of our guidelines to make different stakeholders think of actions that they could take for effective teaming, as well as to evaluate the relevance and clarity of the suggested guidelines. Based on the evaluation results, we refined the guidelines, going from 27 to 24 final guidelines.

6.1 Procedure

The guideline evaluation was divided in two steps: a survey and a think-aloud session. Five experts participated in each step.

Survey. We first conducted a 60-minute survey per expert. At the beginning of the survey, participants were shown a video with instructions for the study. Each participant was shown a scenario describing a pitfall between a human and one of the following LLM agents: (1) personal health assistant, (2) coding assistant, (3) legal assistant, or (4) travel assistant. See the Appendix for more detailed descriptions of the scenarios. Participants were then presented our guidelines. To evaluate the generative power of the guidelines, participants were invited to think of actions that could help prevent the pitfall following these guidelines. The survey guided the participants through all 27 guidelines across 9 different groups of guidelines (one for each thematic grouping e.g., developing shared awareness). For each guideline, participants could consult an example action. While the example action aimed at helping participants ideate, we asked participants to use these examples with caution to avoid anchoring effects [27]. At the end of each group of guidelines, participants evaluated the clarity and relevance of the guidelines for preventing the pitfall. After the survey, we iterated on guidelines that were not considered relevant and improved their clarity.

Think aloud. The second step of the evaluation phase was conducted through a 60-minute think-aloud session per expert. The version of the guidelines used for the think-aloud session was a refined version of the guidelines used in the survey. The objective of the think-aloud sessions was (a) to evaluate the relevance and clarity of the last version of the guidelines and (b) to get a nuanced understanding about experts' perceptions towards the guidelines. A think-aloud format was chosen because it generates richer data than a survey. The procedure for the think-aloud session was the same as for the survey-based study, but instead of asking participants to write down actions, they were asked to orally share their thoughts about potential actions and to reflect on the guidelines.

Participant Recruitment. We recruited 10 participants: 4 specialized in natural language processing (NLP), 3 in NLP and human-computer interaction (HCI), 1 in HCI, 1 in cognitive science, and 1 in management studies. We used purposive sampling, recruiting participants through the authors' networks. Participation was voluntary and was rewarded with a 30 EUR voucher.

Data collection and analysis. Data consists of the ratings of relevance and clarity provided by participants for each guideline group and actions suggested per guideline. Data also includes the transcripts of the think-aloud sessions. We analyzed the transcripts using *reflexive thematic analysis* [12, 23]. We followed the same procedure as in the workshop for data analysis –see section 4.1.

6.2 Results of the Evaluation

We report on the findings from our user study below. We refer to quotes as S-P*j* and TA-P*k*. S stands for *survey*, TA for *think aloud*, P

for *participant* and *j* and *k* are the indices of each quoted participant in the survey and think-aloud studies, respectively. We discard S-P2 from the analysis, due to a possible fatigue effect [104], i.e., from a point on, the quality of their responses dropped noticeably. Unless stated otherwise, when referring to guideline numbers, we refer to the latest 24 guidelines (Figure 1 and Appendix).

Relevance. Figure 2a shows relevance ratings per guideline group. *Survey:* guidelines on Strategy Development were deemed *Not relevant* on two occasions while guidelines on Affect Management were considered *Not relevant* once. Both instances occurred exclusively within the travel assistant scenario. *Think aloud:* guidelines on Coordination and Completion were considered *Not relevant* for the travel scenario. The personal health assistant scenario also raised questions about the applicability of guidelines on Structuring Teamwork, Coordinating, and Affect Management. In contrast, in the coding and legal assistant scenarios all guidelines were considered relevant, with only minor doubts expressed. Relevance ratings revealed a divide between use cases across both studies. Scenarios situated within workplace settings—where humans and LLM agents collaborate in work-related tasks (coding and legal scenarios)—aligned well with our guidelines. However, consumer-facing scenarios—where a human consumer is assisted by an LLM agent (travel and health assistant scenarios)—presented a different picture. The divide was especially evident in the travel assistant scenario: "*Who is the team here? Is there a a customer service team working with the agent? Or are we talking about the customer?*" (TA-P3). Consumer-facing scenarios were *perceived* as one-shot interactions while our guidelines are specifically designed to capture the iterative and ongoing nature of teaming processes. This might indicate that the teaming perspective underlying our guidelines may be most relevant for human-LLM agent teams within workplace settings.

Clarity. Figure 2b shows clarity ratings per guideline group. *Survey:* we identified four ratings indicating *Unclear* guideline groups. Two of those ratings involved guidelines on Strategy Development (S-P3, S-P5), and one involved guidelines on Monitoring and Safeguarding (S-P3). All those ratings were connected to the travel assistant scenario and might, therefore, be the result of the mismatch between the teaming lens and consumer-facing scenarios. Guidelines on Structuring Teamwork and Developing Shared awareness were rated as *Neutral* in clarity in 2 out of 4 and 3 out of 4 cases, respectively. We iterated on those guidelines e.g., *"Adapt the interaction medium to the context"* turned into *"Select the interaction medium based on team needs"* (G2), which helped disambiguate the term *context*. We also changed the order of the guidelines related to Structuring Teamwork (first group in the latest version) and Developing Shared Awareness (second group in the latest version). This better reflects the temporal dependencies among guidelines. *Think aloud:* most of the guidelines were rated as *Clear* or *Very clear*. The Continuous Learning group of guidelines was the only one considered *Unclear*. The number of ratings under the *Neutral* label were also significantly reduced compared to the survey study. Transcripts of the think-aloud session indicate that guidelines on Continuous Learning were found to *"seemingly overlap"* (TA-P3). We, therefore, merge guidelines originally numbered as 23 and 24 into one guideline (G20 in the latest version) where both the human and the LLM agent learn from each other's past interactions and feedback. We also merge guidelines originally numbered as 6 and 7 (G5 in the latest version), and 11 and 12 (G10 in the latest version). While these guidelines were already considered to be clear, participants indicated merging thematically-similar guidelines would help obtain a more concise version of the guidelines.

Experts' reflections on the guidelines. The think-aloud session enabled us to capture experts' reflections on our guidelines. We highlight three main reflections that suggest minor modifications. First, a recurring comment was the potential negative side effects of the suggested guidelines. Guidelines indicating that LLM agents should provide step-by-step explanations or be transparent about data sources were perceived to be appropriate, yet they could also lead to humans overrelying on LLM agents. We, therefore, modify "*Provide explanations for output verification*" to "*Provide information for output verification*" (G10). Second, expert participants suggested G8 (originally directed at the interacting human) might be too demanding for lay people, even if they receive training in AI. We, therefore, assign G8 to the team designer. Third, the desirability of guidelines under Affect Management was challenged. Questions were raised about the desirability of improving perceived competence of LLM agents and providing them with emotional intelligence. Accordingly, we reword G22 from "*to improve perceived competence*" to "*to ensure perceived competence matches LLM agents' capabilities*", and we include "*when appropriate*" in G23.

Coverage. Guidelines were perceived to be comprehensive. However, participants suggested additional considerations. We highlight three. First, the need to embed the guidelines in the current regulatory landscape. Since human-LLM agent teaming is part of a digitization process, regulatory efforts such as the EU's AI Act need to be considered. This will affect what "*you can and cannot do with LLM agents*" (TA-P3). Second, the need to extend the unit of analysis from the team level to higher authorities. This would, for instance, require clarifying the *"responsibility of a government"* (S-P3) when improving the population's AI literacy. Third, the granularity and applicability of the guidelines in practice. Organizations are more granular than what we depicted through our guidelines: "*we have procurement, we have innovation teams, we have unions*" (TA-P2). Similarly, guidelines referring to infrastructure or documentation might face the reality of poor existing practices: "*Maintaining a database of failures would be really useful. But people tend to write really bad documentation.*" (TA-P1). Our guidelines should be seen as a starting point to be transferred and refined in real-world organizations. We leave these considerations for future iterations.

7 DISCUSSION AND FUTURE WORK

We introduced 24 guidelines for human-LLM agent teams organized across five development stages of teamwork. The guidelines move beyond technical design of LLM agents to address multiple stakeholders: LLM agents, their human teammates, team designers and the embedding organization. Our evaluation suggests the guidelines are relevant and clear, finding them applicable to workplace teaming scenarios. While providing a framework for principled human-LLM team design, our guidelines also expose key challenges inherent in current LLM-based systems. We suggest that AAMAS's rich body of research can play an important role in addressing these challenges.

Flexible and dynamic teamwork. Our guidelines call for dynamic systems that support flexible team composition (G1-G4), continuous

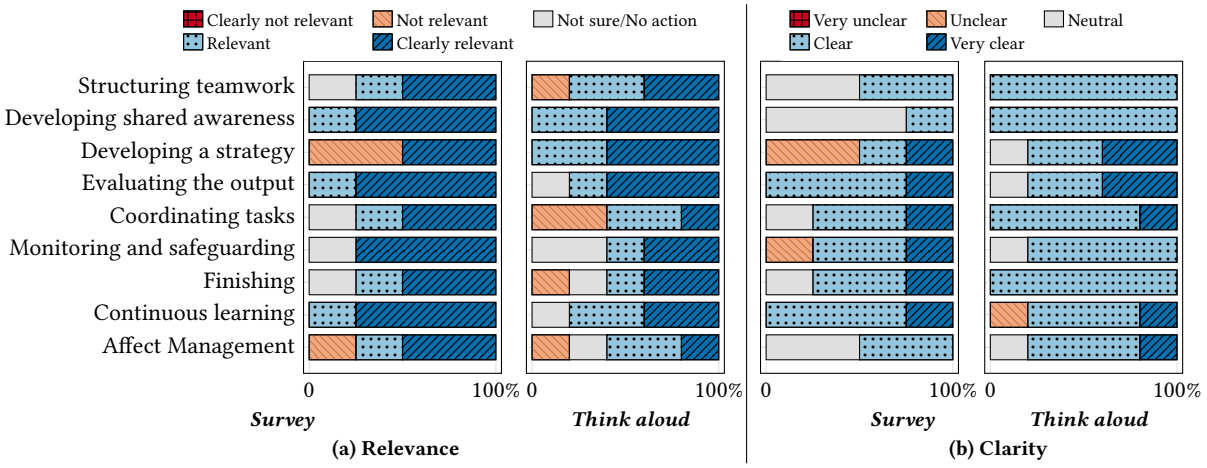

Figure 2: Relevance (a) and clarity (b) ratings from our user study for each guideline group (y axis). The x axis of each plot indicates the percentage of participants (n=4 for *Survey*, n=5 for *Think aloud*).

communication and adaptation (G5-G8) and collaborative strategy development and coordination (G9, G12-G14). Current LLM-based systems, however, rely on predefined workflows and static roles [17, 32, 41, 69], which fundamentally limits their flexibility. Established organization-based MASs methodologies [6, 31, 38, 102] can help bridge this gap, offering robust frameworks to structure teamwork, formalize roles, responsibilities, and norms for human and LLM agents. Research on dynamic team formation [5, 70, 77] and role recognition [59] can inform team active collaboration processes, and the implementation of LLM agents' guidelines using team-oriented prompting strategies (similar to Constitutional AI [7]).

Control and predictability of interactions. Our guidelines demand cotrollable and predictable human-LLM agent teams able to break down and scaffold tasks collaboratively (G9), to ensure transparent execution though team situational awareness, and to course-correct team members that stray from team goals (G15-G16). However, unlike human collaboration, which relies on mutual accountability and trust-building [48, 88], existing LLM agent orchestration uses restricted interaction protocols and prompt-based message exchange [42, 60, 78], making control over error propagation and hallucination difficult. To help address this gap, MAS exogenous regulation mechanisms [2, 13, 18] could be employed to steer the weakly controllable black-box LLM agents. This requires commitment-based interactions [8, 19, 98] for justifiable task execution, and the use of guards, sanctions, and rewards [3, 15, 25] to enforce organizational or team norms. Complementarily, internal LLM reflection techniques [93, 95] can evaluate alignment with team objectives, while specialized agents monitor adherence to team norms. Open challenges remain in creating standardized, team-oriented interaction protocols and ensuring LLM output correctness.

Continuous evolution over time. Our guidelines require human-LLM agent teams engineered for continuous evolution, supporting regular output reassessment (G10-G11), evolving knowledge by providing and learning from feedback (G19-G20), and considering human well-being and (inter)personal team dynamics (G22-G24). Current LLM agents, however, lack adaptability for real-time alignment and continuous learning and feedback [93]. Research on task allocation and coordination [34, 84, 86, 87] and multi-agent reinforcement learning [14, 76] can provide the necessary foundation for continuous adaptation and structured feedback loops. A major open challenge is the emerging evidence of LLM addictive potential [73, 100], which along with human factors such as age, AI literacy or affinity to technology [54, 56, 82] will affect the long-term well-being of humans as part of such teams.

Limitations and Future Work. (1) Expert feedback from our user study confirmed the clarity and relevance of our guidelines, but also suggested potential limitations on their scope and applicability. The teaming lens was perceived as more appropriate in workplace settings than consumer-facing scenarios. Additional work is needed to assess the usability of our guidelines outside workplace settings. (2) Our guidelines represent a starting point and still require evaluation and adaptation in real-world organizational contexts. Future work should ground these guidelines in existing organizational practices and account for context-specific regulatory landscapes, such as the AI Act in the European Union. (3) Our search criteria limited the systematic literature review to academic literature. While some of the reviewed papers discuss findings from industry settings (e.g. [68, 89]), we did not use blog posts of industry releases to inform our guidelines. Future research could extend our guidelines by incorporating those. (4) Throughout our user studies, we involved a total of 25 human-AI teaming experts currently based in academic institutions. While their expertise is varied, the sample size is limited. Future research should involve additional experts, including industry practitioners.

ACKNOWLEDGMENTS

This research was partially supported by Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://www.hybrid-intelligence-centre.nl/, under Grant No. (024.004.022).

REFERENCES

[1] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. 2020. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer* 53, 8 (8 2020), 18–28. https://doi.org/10.1109/MC.2020.2996587

[2] Huib Aldewereld. 2009. Autonomy vs. conformity: an institutional perspective on norms and protocols. *The Knowledge Engineering Review* 24, 4 (2009), 410–411.

[3] Natasha Alechina, Nils Bulling, Mehdi Dastani, and Brian Logan. 2015. Practical Run-Time Norm Enforcement with Bounded Lookahead. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015, Istanbul, Turkey, May 4-8, 2015*, Gerhard Weiss, Pinar Yolum, Rafael H. Bordini, and Edith Elkind (Eds.). ACM, 443–451. http://dl.acm.org/citation.cfm?id=2772937

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[5] Ewa Andrejczuk, Rita Berger, Juan A Rodriguez-Aguilar, Carles Sierra, and Víctor Marín-Puchades. 2018. The composition and formation of effective teams: computer science meets organizational psychology. *The Knowledge Engineering Review* 33 (2018), e17.

[6] Estefania Argente, Vicente Julián, and Vicente J. Botti. 2005. Multi-Agent System Development Based on Organizations. In *Proceedings of the First International Workshop on Coordination and Organisation, CoOrg@COORDINATION 2005, Namur, Belgium, April 23, 2005 (Electronic Notes in Theoretical Computer Science, Vol. 150)*, Guido Boella and Leendert van der Torre (Eds.). Elsevier, 55–71. https://doi.org/10.1016/J.ENTCS.2006.03.005

[7] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[8] Matteo Baldoni, Cristina Baroglio, Federico Capuzzimati, and Roberto Micalizio. 2018. Commitment-based Agent Interaction in JaCaMo+. *Fundam. Informaticae* 159, 1-2 (2018), 1–33. https://doi.org/10.3233/FI-2018-1656

[9] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI, 2019*. AAAI Press, 2429–2437.

[10] Gagan Bansal, Jennifer Wortman Vaughan, Saleema Amershi, Eric Horvitz, Adam Fourney, Hussein Mozannar, Victor Dibia, and Daniel S Weld. 2024. Challenges in human-agent communication. *arXiv preprint arXiv:2412.10380* (2024).

[11] Michael T Brannick, Ashley Prince, Carolyn Prince, and Eduardo Salas. 1995. The measurement of team process. *Human Factors* 37, 3 (1995), 641–651.

[12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (1 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[13] Nils Bulling and Mehdi Dastani. 2016. Norm-based mechanism design. *Artif. Intell.* 239 (2016), 97–142. https://doi.org/10.1016/J.ARTINT.2016.07.001

[14] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Trans. Syst. Man Cybern. Part C* 38, 2 (2008), 156–172. https://doi.org/10.1109/TSMCC.2007.913919

[15] Henrique Lopes Cardoso and Eugénio C. Oliveira. 2009. Adaptive Deterrence Sanctions in a Normative Framework. In *Proceedings of the 2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2009, Milan, Italy, 15-18 September 2009*. IEEE Computer Society, 36–43. https://doi.org/10.1109/WI-IAT.2009.123

[16] Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. Interdependence and trust analysis (ITA): a framework for human–machine team design. *Behaviour & Information Technology* (11 2024), 1–21. https://doi.org/10.1080/0144929X.2024.2431631

[17] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201* (2023).

[18] Amit Chopra, Leendert van der Torre, Harko Verhagen, and Serena Villata. 2018. *Handbook of normative multiagent systems*. College Publications.

[19] Amit K. Chopra, Matteo Baldoni, Samuel Christie, and Munindar P. Singh. 2025. Azorus: Commitments over Protocols for BDI Agents. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025, Detroit, MI, USA, May 19-23, 2025*, Sanmay Das, Ann Nowé, and Yevgeniy Vorobeychik (Eds.). International Foundation for Autonomous Agents and Multiagent Systems / ACM, 490–499. https://doi.org/10.5555/3709347.3743564

[20] Nazli Cila. 2022. Designing Human-Agent Collaborations: Commitment, responsiveness, and support. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18. https://doi.org/10.1145/3491102.3517500

[21] Nazli Cila, Gabriele Ferri, Martijn de Waal, Inte Gloerich, and Tara Karpinski. 2020. The Blockchain and the Commons: Dilemmas in the Design of Local Platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376660

[22] Victoria Clarke and Virginia Braun. 2013. *Successful qualitative research: A practical guide for beginners*. Sage publications ltd. 1–400 pages.

[23] Victoria Clarke and Virginia Braun. 2021. *Thematic analysis: a practical guide*. SAGE Publications Ltd.

[24] Petru Lucian Curseu. 2006. Emergent states in virtual teams: a complex adaptive systems perspective. *Journal of Information Technology* 21, 4 (2006), 249–261.

[25] Davide Dell'Anna, Mehdi Dastani, and Fabiano Dalpiaz. 2020. Runtime revision of sanctions in normative multi-agent systems. *Auton. Agents Multi Agent Syst.* 34, 2 (2020), 43. https://doi.org/10.1007/S10458-020-09465-8

[26] Davide Dell'Anna, Pradeep K Murukannaiah, Bernd Dudzik, Davide Grossi, Catholijn M Jonker, Catharine Oertel, and Pnar Yolum. 2024. Toward a Quality Model for Hybrid Intelligence Teams. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 434–443.

[27] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A Checklist to Combat Cognitive Biases in Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9, 1 (10 2021), 48–59. https://ojs.aaai.org/index.php/HCOMP/article/view/18939

[28] Wen Duan, Christopher Flathmann, Nathan McNeese, Matthew J Scalia, Ruihao Zhang, Jamie Gorman, Guo Freeman, Shiwen Zhou, Allyson Ivy Hauptman, and Xiaoyun Yin. 2025. Trusting Autonomous Teammates in Human-AI Teams - A Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–23. https://doi.org/10.1145/3706598.3713527

[29] Gwendolyn Edgar, Matthew McWilliams, and Matthias Scheutz. 2023. Improving Human-Robot Team Performance with Proactivity and Shared Mental Models. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS, 2023*. ACM, 2322–2324.

[30] Luciano Floridi. 2019. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology* 32, 2 (6 2019). https://doi.org/10.1007/s13347-019-00354-x

[31] Nicoletta Fornara and Charalampos Tampitsikas. 2013. Semantic technologies for open interaction systems. *Artificial Intelligence Review* 39, 1 (2013), 63–79.

[32] Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142* (2023).

[33] Uğur Genç and Himanshu Verma. 2024. Situating Empathy in HCI/CSCW: A Scoping Review. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (11 2024), 1–37. https://doi.org/10.1145/3687052

[34] Athina Georgara, Juan Antonio Rodríguez-Aguilar, and Carles Sierra. 2021. Towards a competence-based approach to allocate teams to tasks. (2021).

[35] Athina Georgara, Juan A Rodriguez Aguilar, and Carles Sierra. 2022. Building contrastive explanations for multi-agent team formation. In *Proceedings of the 21st international conference on autonomous agents and multiagent systems*. 516–524.

[36] Felix Gervits, Dean Thurston, Ravenna Thielstrom, Terry Fong, Quinn Pham, and Matthias Scheutz. 2020. Toward Genuine Robot Teammates: Improving Human-Robot Team Performance Using Robot Shared Mental Models. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2020*. International Foundation for Autonomous Agents and Multiagent Systems, 429–437.

[37] Victoria Groom and Clifford Nass. 2007. Can robots be teammates? *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 8, 3 (11 2007), 483–500. https://doi.org/10.1075/is.8.3.10gro

[38] Davide Grossi, John-Jules Ch Meyer, and Frank Dignum. 2006. Classificatory aspects of counts-as: An analysis in modal logic. *Journal of Logic and Computation* 16, 5 (2006), 613–643.

[39] J Richard Hackman. 2002. *Leading teams: Setting the stage for great performances*. Harvard Business Press.

[40] Andreas T Hirblinger. 2022. When Mediators Need Machines (and Vice Versa): Towards a Research Agenda on Hybrid Peacemaking Intelligence. *International Negotiation* 28, 1 (2022), 94–125.

[41] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. International Conference on Learning Representations, ICLR.

[42] William Hunt, Toby Godfrey, and Mohammad D Soorati. 2024. Conversational Language Models for Human-in-the-Loop Multi-Robot Coordination. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 2809–2811.

[43] Daniel R Ilgen, John R Hollenbeck, Michael Johnson, and Dustin Jundt. 2005. Teams in organizations: From input-process-output models to IMOI models. *Annu. Rev. Psychol.* 56 (2005), 517–543.

[44] Daniel R Ilgen, John R Hollenbeck, Michael Johnson, and Dustin Jundt. 2005. Teams in organizations: From input-process-output models to IMOI models. *Annual Review of Psychology* 56 (2005), 517–543.

[45] Matthew Johnson, Jeffrey M Bradshaw, Paul J Feltovich, Catholijn M Jonker, M Birna Van Riemsdijk, and Maarten Sierhuis. 2014. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction* 3, 1 (2014), 43–69.

[46] Matthew Johnson and Alonso H. Vera. 2019. No AI Is an Island: The Case for Teaming Intelligence. *AI Magazine* 40, 1 (3 2019), 16–28. https://doi.org/10.1609/aimag.v40i1.2842

[47] Mladan Jovanovic and Mia Schmitz. 2022. Explainability as a User Requirement for Artificial Intelligence Systems. *Computer* 55, 2 (2022), 90–94.

[48] Jon R Katzenbach and Douglas K Smith. 2005. The discipline of teams. *Harvard business review* 83, 7 (2005), 162.

[49] Gary Klein, Paul J Feltovich, Jeffrey M Bradshaw, and David D Woods. 2005. Common ground and coordination in joint activity. *Organizational simulation* 53 (2005), 139–184.

[50] Gary Klein, David D. Woods, Jeffrey M. Bradshaw, Robert R. Hoffman, and Paul J. Feltovich. 2004. Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. *IEEE Intelligent Systems* 19, 6 (2004), 91–95.

[51] Richard Klimoski and Susan Mohammed. 1994. Team mental model: Construct or metaphor? *Journal of management* 20, 2 (1994), 403–437.

[52] Noam Kolt. 2025. Governing AI agents. *arXiv preprint arXiv:2501.07913* (2025).

[53] E. S. Kox, Jose H. Kerstholt, T. F. Hueting, and P. W. de Vries. 2022. Trust Repair in Human-Agent Teams: The Effectiveness of Explanations and Expressing Regret. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 1944–1946.

[54] Max F. Kramer, Jana Schaich Borg, Vincent Conitzer, and Walter Sinnott-Armstrong. 2018. When Do People Want AI to Make Decisions?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 204–209. https://doi.org/10.1145/3278721.3278752

[55] Kirill Krinkin and Yulia A. Shichkina. 2022. Cognitive Architecture for Co-evolutionary Hybrid Intelligence. In *Proceedings of the 15th International Conference on Artificial General Intelligence, AGI 2022 (Lecture Notes in Computer Science, Vol. 13539)*. 293–303.

[56] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. 2022. "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–28. https://doi.org/10.1145/3491102.3517527

[57] Lindsay Larson and Leslie A DeChurch. 2020. Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *The leadership quarterly* 31, 1 (2020), 101377.

[58] Li Lin, Yu Wang, Yayu Ping, Jian Gao, Zongbo Wang, and Shouyu Wang. 2025. Design Guidelines for Human-Generative AI Interaction. 223–239. https://doi.org/10.1007/978-3-031-93718-7{_}15

[59] Linus J Luotsinen and Ladislau Bölöni. 2008. Role-based teamwork activity recognition in observations of embodied agent actions. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*. 567–574.

[60] Zhao Mandi, Shreeya Jain, and Shuran Song. 2024. Roco: Dialectic multi-robot collaboration with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 286–299.

[61] Michelle A. Marks, John E. Mathieu, and Stephen J. Zaccaro. 2001. A Temporally Based Framework and Taxonomy of Team Processes. *The Academy of Management Review* 26, 3 (7 2001), 356. https://doi.org/10.2307/259182

[62] Michelle A Marks, John E Mathieu, and Stephen J Zaccaro. 2001. A temporally based framework and taxonomy of team processes. *Academy of management review* 26, 3 (2001), 356–376.

[63] Joseph Edward McGrath. 1984. *Groups: Interaction and performance*. Vol. 14. Prentice-Hall Englewood Cliffs, NJ.

[64] Joseph E McGrath, Holly Arrow, and Jennifer L Berdahl. 2014. The study of groups: Past, present, and future. In *Personality and Social Psychology at the Interface*. Psychology Press, 95–105.

[65] Siddharth Mehrotra. 2021. Modelling Trust in Human-AI Interaction. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, AAMAS, 2021*. ACM, 1826–1828.

[66] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors* 58, 3 (2016), 401–415.

[67] Christopher Myers, Jerry Ball, Nancy Cooke, Mary Freiman, Michelle Caisse, Stuart Rodgers, Mustafa Demir, and Nathan McNeese. 2018. Autonomous intelligent agents for team training. *IEEE Intelligent Systems* 34, 2 (2018), 3–14.

[68] Suchismita Naik, Austin L. Toombs, Ph.D. Snellinger, Amanda, Scott Saponas, and Amanda K Hall. 2025. Designing with Multi-Agent Generative AI: Insights from Industry Early Adopters. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 1961–1972. https://doi.org/10.1145/3715336.3735823

[69] Nathalia Nascimento, Paulo Alencar, and Donald Cowan. 2023. Self-adaptive large language model (llm)-based multiagent systems. In *2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*. IEEE, 104–109.

[70] Dung Nguyen, Phuoc Nguyen, Svetha Venkatesh, and Truyen Tran. 2022. Learning to Transfer Role Assignment Across Team Sizes. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 963–971.

[71] OpenAI. 2025. A practical guide to building agents.

[72] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2022. Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 64, 5 (8 2022), 904–938. https://doi.org/10.1177/0018720820960865

[73] Nizan Geslevich Packin and Karni Chagal-Feferkorn. 2024. This is not a game: The addictive allure of digital companions. *Seattle UL Rev.* 48 (2024), 693.

[74] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30, 3 (5 2000), 286–297. https://doi.org/10.1109/3468.844354

[75] David V. Pynadath, Nikolos Gurney, Sarah Kenny, Rajay Kumar, Stacy C. Marsella, Haley Matuszak, Hala Mostafa, Pedro Sequeira, Volkan Ustun, and Peggy Wu. 2023. Effectiveness of Teamwork-Level Interventions through Decision-Theoretic Reasoning in a Minecraft Search-and-Rescue Task. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023*. ACM, 2334–2336.

[76] Roxana Radulescu. 2024. The World is a Multi-Objective Multi-Agent System: Now What?. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024) (Frontiers in Artificial Intelligence and Applications, Vol. 392)*, Ulle Endriss, Francisco S. Melo, Kerstin Bach, Alberto José Bugarín Diz, Jose Maria Alonso-Moral, Senén Barro, and Fredrik Heintz (Eds.). IOS Press, 32–38. https://doi.org/10.3233/FAIA240464

[77] Pranav Rajbhandari, Prithviraj Dasgupta, and Donald Sofge. 2025. Transformer Guided Coevolution: Improved Team Formation in Multiagent Adversarial Games. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. 2720–2722.

[78] Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. 2023. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928* (2023).

[79] Fardin Saad, Pradeep K Murukannaiah, and Munindar P Singh. 2025. Gricean Norms as a Basis for Effective Collaboration. *arXiv preprint arXiv:2503.14484* (2025).

[80] Eduardo Salas, C Shawn Burke, and Janis A Cannon-Bowers. 2000. Teamwork: emerging principles. *International Journal of Management Reviews* 2, 4 (2000), 339–356.

[81] Eduardo Salas, Terry L Dickinson, Sharolyn A Converse, and Scott I Tannenbaum. 1992. *Toward an understanding of team performance and training*. Ablex Publishing.

[82] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. (5 2022). https://doi.org/10.1145/3531146.3533218

[83] Isabella Seeber, Eva Bittner, Robert O. Briggs, Triparna de Vreede, Gert-Jan de Vreede, Aaron Elkins, Ronald Maier, Alexander B. Merz, Sarah Oeste-Reiß, Nils Randrup, Gerhard Schwabe, and Matthias Söllner. 2020. Machines as teammates: A research agenda on AI in team collaboration. *Information & Management* 57, 2 (3 2020), 103174. https://doi.org/10.1016/j.im.2019.103174

[84] Ameesh Shah, Niklas Lauffer, Thomas Chen, Nikhil Pitta, and Sanjit A Seshia. 2025. Learning Symbolic Task Decompositions for Multi-Agent Teams. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. 1904–1913.

[85] Dominik Siemon. 2022. Elaborating Team Roles for Artificial Intelligence-based Teammates in Human-AI Collaboration. *Group Decision and Negotiation* 31, 5 (10 2022), 871–912. https://doi.org/10.1007/s10726-022-09792-z

[86] Susannah Soon, Adrian Pearce, and Max Noble. 2003. Modelling the collaborative mission planning process using dynamic teamwork structures. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*. 1124–1125.

[87] Susannah Soon, Adrian Pearce, and Max Noble. 2004. Adaptive teamwork coordination using graph matching over hierarchical intentional structures. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, Vol. 1. IEEE Computer Society, 294–301.

[88] Virginia R Stewart, Deirdre G Snyder, and Chia-Yu Kou. 2023. We hold ourselves accountable: A relational view of team accountability. *Journal of Business Ethics* 183, 3 (2023), 691–712.

[89] Macy Takaffoli, Sijia Li, and Ville Makela. 2024. Generative AI in User Experience Design and Research: How Do UX Practitioners, Teams, and Companies Use GenAI in Industry?. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 1579–1593. https://doi.org/10.1145/3643834.3660720

[90] Emma Van Zoelen, Tina Mioch, Mani Tajaddini, Christian Fleiner, Stefani Tsaneva, Pietro Camin, Thiago S. Gouvêa, Kim Baraka, Maaike H. T. De Boer, and Mark A. Neerincx. 2023. Developing Team Design Patterns for Hybrid Intelligence Systems. https://doi.org/10.3233/FAIA230071

[91] Emma Van Zoelen, Tina Mioch, Mani Tajaddini, Christian Fleiner, Stefani Tsaneva, Pietro Camin, Thiago S. Gouvêa, Kim Baraka, Maaike H. T. De Boer, and Mark A. Neerincx. 2023. Developing Team Design Patterns for Hybrid Intelligence Systems. https://doi.org/10.3233/FAIA230071

[92] Giovanna Varni, André-Marie Pez, and Maurizio Mancini. 2021. Get Together in the Middle-earth: a First Step Towards Hybrid Intelligence Systems. In *Proceedings of the 2021 International Conference on Multimodal Interaction, ICMI 2021, Companion Publication*. ACM, 249–253.

[93] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 186345.

[94] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. ACM, 997–1005.

[95] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300* (2023).

[96] Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–22. https://doi.org/10.1145/3613904.3642466

[97] Chathurika S Wickramasinghe, Daniel L Marino, Javier Grandio, and Milos Manic. 2020. Trustworthy AI development guidelines for human system interaction. In *2020 13th International Conference on Human System Interaction (HSI)*. IEEE, 130–136.

[98] Michael Winikoff. 2007. Implementing commitment-based interactions. In *6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2007), Honolulu, Hawaii, USA, May 14-18, 2007*, Edmund H. Durfee, Makoto Yokoo, Michael N. Huhns, and Onn Shehory (Eds.). IFAAMAS, 128. https://doi.org/10.1145/1329125.1329283

[99] Zihan Yan and Yaohong Xiang. 2025. Social life simulation for non-cognitive skills learning. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–44.

[100] Ala Yankouskaya, Magnus Liebherr, and Raian Ali. 2025. Can ChatGPT be addictive? A call to examine the shift from support to dependence in AI conversational large language models. *Human-Centric Intelligent Systems* (2025), 1–13.

[101] Mireia Yurrita, Himanshu Verma, Agathe Balayn, Ujwal Gadiraju, Sylvia C Pont, and Alessandro Bozzon. 2025. Towards Effective Human Intervention in Algorithmic Decision-Making: Understanding the Effect of Decision-Makers' Configuration on Decision-Subjects' Fairness Perceptions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.

[102] Franco Zambonelli, Nicholas R. Jennings, and Michael J. Wooldridge. 2001. Organisational Rules as an Abstraction for the Analysis and Design of Multi-Agent Systems. *Int. J. Softw. Eng. Knowl. Eng.* 11, 3 (2001), 303–328. https://doi.org/10.1142/S0218194001000505

[103] Qiaoning Zhang, Matthew L. Lee, and Scott A. Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI 2022*. 114:1–114:28.

[104] Ying Zhang, Xianghua Ding, and Ning Gu. 2018. Understanding fatigue and its impact in crowdsourcing. In *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*. IEEE, 57–62.

[105] Zhongyi Zhou, Jing Jin, Vrushank Phadnis, Xiuxiu Yuan, Jun Jiang, Xun Qian, Kristen Wright, Mark Sherwood, Jason Mayes, Jingtao Zhou, et al. 2025. InstructPipe: Generating Visual Blocks Pipelines with Human Instructions and LLMs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.