Toward a Quality Model for Hybrid Intelligence Teams

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

Bernd Dudzik
Delft University of Technology
Delft, The Netherlands
b.j.w.dudzik@tudelft.nl

Davide Grossi
University of Groningen
Groningen, The Netherlands
d.grossi@rug.nl

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

Catharine Oertel
Delft University of Technology
Delft, The Netherlands
c.r.m.m.oertel@tudelft.nl

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

ABSTRACT

Hybrid Intelligence (HI) is an emerging paradigm in which artificial intelligence (AI) augments human intelligence. The current literature lacks systematic models that guide the design and evaluation of HI systems. Further, discussions around HI primarily focus on technology, neglecting the holistic human-AI ensemble. In this paper, we take the initial steps toward the development of a quality model for characterizing and evaluating HI systems from a human-AI teams perspective. We conducted a study investigating the adequacy of properties commonly associated with effective human teams to describe HI. Our study, featuring the insights of 50 HI researchers, shows that various human team properties, including boundedness, interdependence, competency, purposefulness, initiative, normativity, and effectiveness, are important for HI systems. Our study also reveals limitations in applying certain human team properties, such as coaching, rewards, and recognition, to HI systems due to the inherent human-AI asymmetry.

KEYWORDS

Hybrid Intelligence; Quality model; Human-agent teamwork; Sociotechnical systems; Team Diagnostic Survey

ACM Reference Format:
Davide Dell'Anna, Pradeep K. Murukannaiah, Bernd Dudzik, Davide Grossi, Catholijn M. Jonker, Catharine Oertel, and Pınar Yolum. 2024. Toward a Quality Model for Hybrid Intelligence Teams. In *Proc.of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 10 pages.

1 INTRODUCTION

Artificial Intelligence (AI) permeates many aspects of our daily lives. Literature suggests that AI should work synergistically with humans instead of replacing them to benefit individuals and society [1, 70]. The concept of *Hybrid Intelligence* (HI) [1, 15], in particular, argues for a combination of human and artificial intelligence instead of their isolated operations. Various interpretations of HI exist in the literature, including: HI as an emergent property of human-machine interactions [7, 60, 66], HI as a form of human-in-the-loop or AI-in-the-loop decision making [48, 68], HI as a type of collective intelligence [7], and HI as a design paradigm [8, 45, 74, 81].

Despite a wide array of interpretations, there is a lack of a quality model [77] that guides the systematic development and evaluation of HI systems [39]. Quality models (e.g., [31]) are conceptual representations of quality characteristics of a product, e.g., a software or a system (in our case a HI system). Quality models play a critical role in assuring product quality, i.e., the degree to which a product satisfies stated and implied stakeholder needs [77].

Existing works on HI adopt a technology-centric perspective [14, 26, 61, 82], delineating requirements and attributes of AI components but neglecting the system-level view that underpins concepts such as synergy and interdependence. Similarly, guidelines and models for quality AI systems (e.g., [17, 20, 23, 69]) mainly address the engineering of AI components. In contrast, we emphasize the human-AI *system* and adopt a team-oriented approach, framing HI systems as human-AI teams (or HI teams) [1, 25, 67, 80]. The teaming perspective allows us to explore human- and system-centric dimensions of HI and go beyond existing technology-oriented views.

We develop an initial quality model for HI teams. We answer two research questions, which help in characterizing two key elements of quality models: quality *attributes* and quality *measures* [57, 77].

RQ1 (Adequacy) To what extent are the properties of human teams adequate to characterize HI teams?
RQ2 (Effectiveness) Which measures of effectiveness of human teams are also important for HI teams?

We design and conduct an empirical study to answer these research questions. Through group discussions and a hands-on exercise, 50 HI researchers evaluated various properties and effectiveness measures derived from the *Team Diagnostic Survey* [76], a well-known instrument employed in practice to diagnose the strengths and weaknesses of human teams.

Proc.of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Our findings show that HI experts consider several properties of effective human teams important for HI teams. Some of these properties are well-understood and directly applicable to characterize HI teams. These encompass the concepts of well-defined and interdependent teams, appropriate team size, members diversity, clear meaningful mission, member autonomy and initiative, and team norms. Further, our study indicates that seven measures of human teams effectiveness, encompassing task performance, quality of group processes, and member satisfaction, are deemed important criteria for evaluating also HI teams effectiveness.

Contribution. HI is an emerging field, and understandably, there is a lack of construct clarity [71] on HI. Our work helps improve the construct clarity on HI by offering a series of observations and recommendations derived from a systematic study involving HI researchers. Further, we show how the insights from our study can be used to create a quality model, consisting of quality attributes and measures, for HI teams. Such a model can help analyze existing HI teams by diagnosing missing attributes or performance problems, useful to improve HI teams over time. The model can also enable quality-driven design and development of HI teams from the onset.

Organization. Section 2 discusses related work. Section 3 provides a background on team diagnostic survey. Section 4 describes our study. Section 5 addresses RQ1 and RQ2. Section 6 provides recommendations for a quality model. Section 7 concludes the paper.

2 RELATED WORK

The ideas underlying the concept of HI, sometimes referred to as *human-machine intelligence* or *human-machine symbiosis* [27, 34, 39] root back to the 50s [16, 44]. The term HI was first used in the late 70s in contexts of cybernetics [46] and ergonomics [10]. Nearly four decades later, the concept of HI has found applications in education [51], medicine [42], and computer vision [82].

Various taxonomies organize knowledge around HI and provide a framework for their design and evaluation [26, 48]. Dellerman *et al.* [14] characterize HI via four dimensions: task characteristics, learning paradigm, human-AI interaction, AI-human interaction. These dimensions are strongly (AI) task-oriented, e.g., they identify four task categories: recognition, prediction, reasoning and action. Pescetelli [61] indicates that HI has various levels related to the algorithm's role, e.g., the AI being an assistant, a peer, a facilitator, or a system-level operator. Zschech *et al.* [82] define design principles for computer vision-based HI, focusing on the AI capabilities.

A similar focus on technology appears in recent initiatives delineating guidelines for ensuring AI systems quality (e.g., [17, 20, 69]). Kuwajima *et al.* [41] integrate EU Guidelines for Trustworthy AI [9] into the ISO25010 [31] standard for system and software quality models. They propose new quality characteristics, such as controllability and explainability. Habibullah *et al.* [23] identify quality requirements for machine learning, such as transparency and explainability. However, the focus on AI components of current literature often downplays the crucial concepts for HI, such as collaboration, synergy, interdependence, and relationship between human and AI agents [1, 21, 79]. This motivates the human-agent team orientation of our work, complementing the existing literature.

There is extensive literature on teamwork (e.g., [3, 4, 55, 72]) and human-AI (including, human-robot and human-agent) teams [36, 59, 66]. Johnson *et al.* [33] provide a design-time approach for handling interdependence between actors of a team that can be translated into control algorithms. Bansan *et al.* [2] study how adaptation of agents at run-time affect the interactions between humans and agents in performing tasks. Zhang *et al.* [80] investigate how complementary expertise between humans and agents play a role in creating team trust, and Kox *et al.* [38] develop strategies for agents to repair trust in human-AI teams. Wang *et al.* [78] design agents that provide explanations for their decisions to establish trust from humans in teams. Georgara *et al.* [18] develop explanation algorithms to clarify why certain teams can be formed with team formation algorithms and others cannot. Gervits *et al.* [19] make use of shared mental models to improve performance of human-robot teams. Paynadath *et al.* [62] design an agent that considers team-level properties when making intervention to a human team.

The works above vastly contribute to different aspects of human-agent teams. However, there is still a lack of studies that pinpoint desired properties that HI teams should exhibit, and instruments that capture these properties into quality models with which teams can be analyzed. As a step in this direction, we start from a human team model and investigate its applicability to HI teams [63].

3 TEAM DIAGNOSTIC SURVEY

Human teams have been studied from many angles [22, 50]. Broadly, a team involves two or more *members* working *interdependently* toward a *common goal* [64] so that team activities result in *collective* success or failure, requiring accountability from all members.

IMO is a well-known conceptual model for teamwork [30, 49]. In IMO, the I represents *team inputs* (team composition, tasks complexity, members' differences), which are converted by *team mediators* (M) into *team outputs* (O)—team's outcomes such as team viability, individual learning, development and satisfaction [24]. Mediators include team processes [6, 47] and emergent states [11, 30, 37]. The former refers to developing and adapting the team's purpose, strategy, structure, mutual monitoring and coordination, affect and conflict management. The latter refers to cognitive qualities of the team, like trust and shared mental models.

Many human team design and assessment tools are inspired from the IMO model (see [73] for a systematic review). The Team Diagnostic Survey (TDS) [76] is one such instrument. Table 1 provides an overview of the TDS. The TDS measures six conditions of effective teams' design [76], divided into *The Essentials* and *The Enablers*. The Essentials are the main conditions that result in a sturdy platform for an effective team. The Enablers are conditions that accelerate how fast teams grow into excellent performers. The TDS also measures three *Key Task Processes* that emerge from the six conditions. These are meant to capture how well the team members are working together considering the extents of their capabilities.

The TDS defines seven *measures of effectiveness* of teams. Three relate to *task performance*, including the satisfaction of the team's users with the quality, quantity, and timeliness of the team's work (*E1*, *Users satisfaction*), the appropriateness of the set of choices members make about how to carry out the work (*E2*, *Strategy appropriateness*), and whether the team brings ample and appropriate talent to bear on the work (*E3*, *Knowledge and skills*). The fourth

Table 1: Overview of the *properties* composing the Team Diagnostic Survey and examples of statements [76] characterizing each property. *(R)* stands for *Reverse* and indicates that the statement provides the opposite description of the property.

Category	Property group	Property	Example of statement about the property
Essentials	Real Team	Bounded	*Team membership is quite clear—every member can identify exactly who is and isn't on the team.*
		Stable	*The team is quite stable, with few changes in membership.*
		Interdependent	*Members of the team have to depend heavily on one another to get the team's work done.*
	Compelling Direction	Clear	*There is great uncertainty and ambiguity about what the team is supposed to accomplish. (R)*
		Challenging	*The team's purposes are so challenging that members have to stretch to accomplish them.*
		Consequential	*The team's purposes are of great consequence for those served by the team.*
	Right People	Diversity	*Members of the team are too dissimilar to work together well. (R)*
		Skills	*Members of the team have more than enough talent and experience for the kind of work that the team does.*
Enablers	Sound Structure	Whole Task	*The team does a whole, identifiable piece of work.*
		Autonomy/Judgment	*The team work requires the members to make many "judgment calls" and take initiative as they carry it out.*
		Knowledge of Results	*Carrying out the team's task automatically generates trustworthy indicators of how well the members are doing.*
		Team Size	*The team is just the right size to accomplish its purposes.*
		Team Norms	*It is clear what is—and what is not—acceptable member behavior in the team.*
	Supportive Context	Rewards/Recognition	*Excellent team performance is rewarded.*
		Information	*It is easy for the team to get any data or forecasts that members need to do their work.*
		Education/Consultation	*The team members do not receive adequate training for the work they have to do. (R)*
		Material Resources	*The team members can readily obtain all the material resources that they need for their work.*
	Coaching	Availability	*The team has readily available expert "coaches" who can help it learn from its successes and mistakes.*
		Helpfulness	*Coaches know how and when to intervene.*
Key Task Processes		Effort	*Members demonstrate their commitment to the team by putting in extra time and effort to help it succeed.*
		Strategy	*The team can come up with innovative ways of proceeding with the work.*
		Knowledge and Skill	*Members of the team actively share their special knowledge and expertise with one another.*

measure concerns the quality of group processes and team interactions (*E4*, *Quality of group processes*). The remaining measures concern *member satisfaction*, including satisfaction with the relations between members (*E5*, *Satisfaction with relations*), team members' opportunities to grow and learn over time (*E6*, *Growth opportunities*), and the general team member' satisfaction (*E7*, *General satisfaction*).

The extensive validation of the TDS [73, 76] has shown that the TDS provides a good trade-off between simplicity, abstractness and coverage of concepts concerning human teams.

4 USER STUDY

We describe the design and execution of an empirical study to answer our research questions. Our study seeks to elicit expert knowledge by facilitating discussions among the participants. The discussions yield diverse perspectives and promote iterative refinement of the ideas before the participants answer a questionnaire.

A key objective for our study is to foster a human-centered discussion of human-AI teams as opposed to a purely technology-oriented discussions. Thus, we employed the Team Diagnostic Survey (TDS) as a starting point for the discussions on relevant properties of HI teams. In addition to practical relevance, the TDS is strongly human-oriented. This encouraged discussions on the application of human aspects of teamwork to HI teams.

4.1 Study Design

Figure 1 illustrates the key phases of our study. In the preparation phase, we select four HI teams to provide contexts for discussions, and adapt the TDS questions. In the plenary introduction, we introduce the selected HI teams and the TDS questions to the participants. The participants discuss the TDS questions with respect to the HI teams in the group discussion phase. Finally, the participants answer the TDS questions, individually, via an online survey.

4.1.1 HI Teams. The TDS is used to evaluate a real team, which provides the necessary context to answer the TDS questions. To create such contexts for our study, we formulate four HI teams.

The Warehouse team consists of humans and robots collaborating to manage a warehouse by, e.g., restocking shelves, processing orders, and counting inventory.

The Entertainment team consists of a person and an AI recommender (such as Netflix), interacting to find shows of interest to the person. Interactions take the form of the recommender advising the human, who responds with implicit feedback through selection or rejection of recommendations.

The Research team consists of a researcher and an AI assistant (such as Elicit), interacting, via natural language and a GUI, to conduct research activities such as a literature review or writing a research proposal [5, 56].

The Shepherd team consists of a shepherd and one or more shepherd dogs. Together they take care—guarding, moving, managing and controlling—of sheep.

Teams similar to the ones above exist in practice. This helps ground our study in realistic contexts. Whereas the first three teams are human-AI teams, the last team is a human-dog team, inspired by the ideas of using inter-species collaboration and interdependence as a metaphor for human-machine interaction [28, 43].

Besides being realistic, there is a large variety among these teams in terms of, e.g., application domain, team size and structure, types of AI (e.g., embodied vs. software agents), and types of interactions (simple clicks vs. natural language vs. multimodal). Having a variety of teams helps in eliciting a variety of perspectives.

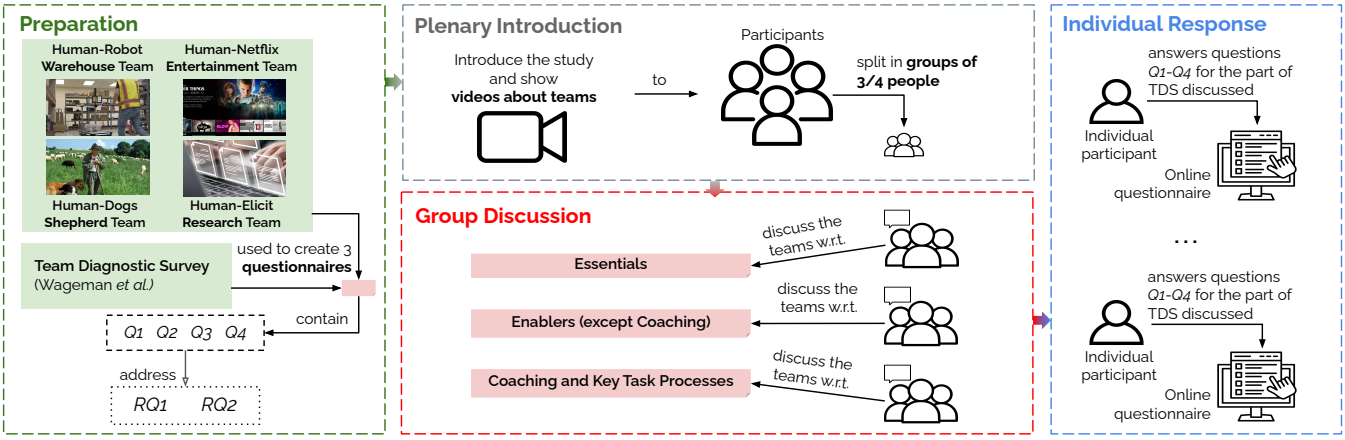

Figure 1: An overview of the key phases of our empirical study.

4.1.2 TDS Questionnaires. We formulate three questionnaires based on the TDS—one for each category of properties in Table 1. We divide the TDS into three parts to reduce the workload for the participants (each completing a different part of the TDS) with an estimated completion time of 45 minutes. To allocate a similar workload to all participants, we move the *Coaching* property group from the *Enablers* to the *Key Task Processes* questionnaire. Each property is illustrated via a number of statements extracted from [76]. The full list of 61 statements used in our questionnaires can be found in our online appendix [13] (Table 1 reports examples).

The questionnaires contain three questions for each property. (*Q1*) How well each team reflects the statements concerning the property (5-point Likert scale). (*Q2*) Explain your evaluation (free text), clarifying what aspects had been considered, whether a team had or lacked an important characteristic or process, whether the statements were difficult to evaluate, and whether the evaluation would have changed if a team exhibited a certain property or worked differently. (*Q3*) Whether the statements are important for HI systems (yes/no). Finally, each questionnaire contains a question (*Q4*) Indicate the importance (5-point Likert scale) of the human teams' effectiveness measures (*E1–E7* in Section 3) for HI systems.

4.1.3 Plenary Introduction. The aim of this phase is to provide a good description of the HI teams and the questions that the participants need to answer. After an introduction about the study, the participants are shown short videos about the four HI teams, and the participants have an opportunity to ask for clarifications.

The participants are asked to form groups of three or four people. Each group randomly receives one of the three types of questionnaires, ensuring a balanced distribution, and a printed version of the questionnaires and team descriptions (see online appendix [13]).

4.1.4 Group Discussion: The aim of this phase is to stimulate the participants' thoughts on both the HI teams and the TDS questions assigned to them. The participants are asked to go through each property and each statement (see Table 1) one by one, and are encouraged to discuss concepts, example behaviors, and missing points. However, the participants are neither asked to answer the questionnaire as a group nor required to reach unanimous decisions.

4.1.5 Individual Response. After the group discussion, participants are asked to individually fill out an online questionnaire on the TDS questions they had group discussions. We opted for individual participant responses to enable the expression of individual opinions and to provide sufficient time to elaborate the responses.

4.2 Study Execution

The study was approved by Utrecht University Ethics assessment Committee. The participants were all researchers actively engaged in projects related to Hybrid Intelligence. All participants had at least an MSc degree. Their ages ranged from 23 to 65. All the participants that took part in the Individual Response phase completed a consent form. Participants were not paid for their participation.

The Plenary introduction (approx. 30 minutes) and Group discussion (approx. 2.5 hours) phases of our study took place in May 2023. A total of 50 participants joined the Plenary introduction phase, which resulted in 15 groups in the Group discussion phase. The Individual response phase took place asynchronously during May and June 2023. 15 participants completed the Individual Response phase, broadly covering the outcomes of the group discussions. Thus, questions *Q1-Q3* were answered by five participants for each questionnaire. Question *Q4* was answered by 15 participants.

5 RESULTS

We analyze the study participants' individual responses to answer the research questions posed earlier. To answer RQ1, we analyze the responses to questions *Q1-Q3*. To answer RQ2, we analyze the participants responses to *Q4*. For each RQ, we mark main observations with a 👍 to indicate properties of human teams that could be integrated into a quality model for HI teams. In contrast, observations marked with a 👎 point toward potential challenges in extending the properties of human teams to HI teams.

5.1 RQ1 (Adequacy)

To answer RQ1, we focus on the properties of human teams that are considered in the TDS. For each property, Table 2 shows, first, if the participants found the property important (*Q3*), second if the property was *well-understood* for HI teams (i.e., no difficulty

Table 2: The number of participants considering a property important for HI (*Q3*), whether (yes/no) the participants reported difficulty in understanding a property (*Q2*), and the scores (mean ± SD) assigned by the participants to the four HI teams (*Q1*).

	Property	Is important for HI (Q3)	Is well understood for HI teams (Q2)	Team's scores (Q1)			
				Warehouse	Entertainment	Research	Shepherd
Essentials	Real Team - Bounded	5 (100%)	Yes	3.2 ± 1.47	3.4 ± 1.36	3.2 ± 1.6	4 ± 1.55
	Real Team - Stable	4 (80%)	Yes	2.8 ± 0.75	3.4 ± 1.02	4 ± 0.89	3.8 ± 0.75
	Real Team - Interdependent	5 (100%)	Yes	2.8 ± 1.6	2.6 ± 0.8	2 ± 0.89	4.4 ± 0.8
	Compelling Direction - Clear	5 (100%)	Yes	4.2 ± 0.75	2.8 ± 1.6	3.2 ± 1.17	4 ± 1.55
	Compelling Direction - Challenging	4 (80%)	No	3 ± 1.41	3.2 ± 0.4	3 ± 0.63	3.4 ± 1.02
	Compelling Direction - Consequential	5 (100%)	Yes	2.4 ± 1.5	3.6 ± 1.2	3.2 ± 1.17	3.8 ± 0.98
	Right People - Diversity	5 (100%)	Yes	3.6 ± 1.36	2.4 ± 1.2	2.6 ± 1.02	3.6 ± 1.2
	Right People - Skills	4 (80%)	No	2.8 ± 1.17	3 ± 1.41	2.2 ± 1.17	3.4 ± 1.62
Enablers	Sound Structure - Whole Task	3 (60%)	No	3.2 ± 1.17	2.8 ± 1.17	3.6 ± 0.49	4.2 ± 1.17
	Sound Structure - Autonomy and judgment	5 (100%)	Yes	3.6 ± 1.36	2.8 ± 0.4	3.6 ± 1.02	3.8 ± 1.6
	Sound Structure - Knowledge of results	4 (80%)	No	2.4 ± 0.49	2.4 ± 1.02	2.4 ± 0.49	4.6 ± 0.49
	Sound Structure - Team Size	5 (100%)	Yes	2.8 ± 0.75	4 ± 1.1	4.2 ± 0.75	3.6 ± 0.8
	Sound Structure - Team Norms	5 (100%)	Yes	3.4 ± 0.8	2.6 ± 0.8	3.2 ± 0.4	4.4 ± 0.8
	Supportive Context - Rewards and recognition	3 (60%)	No	2.2 ± 1.17	2.8 ± 1.17	2.8 ± 0.4	3.6 ± 0.49
	Supportive Context - Information	5 (100%)	No	4 ± 0.63	2.2 ± 0.98	3.4 ± 1.36	2.8 ± 0.4
	Supportive Context - Education and consultation	4 (80%)	No	3.6 ± 0.8	1.8 ± 0.4	3 ± 0.89	3.8 ± 0.98
	Supportive Context - Material Resources	5 (100%)	No	3.8 ± 1.17	2.2 ± 0.98	3 ± 1.79	3.2 ± 0.75
	Coaching - Availability	3 (60%)	No	3.4 ± 0.49	2 ± 1.55	2 ± 1.1	3.4 ± 0.49
	Coaching - Helpfulness	3 (60%)	No	2.8 ± 1.17	1.8 ± 0.98	2 ± 0.89	3.4 ± 0.8
Key Task Processes	Effort	3 (60%)	No	3.2 ± 1.47	2 ± 0.63	2.8 ± 1.17	3.8 ± 1.47
	Strategy	5 (100%)	No	3 ± 1.1	2 ± 1.1	2.4 ± 1.02	3.6 ± 1.36
	Knowledge and skills	5 (100%)	No	2.8 ± 0.75	2 ± 0.63	3 ± 0.89	3.2 ± 1.47

was reported in *Q2* on the understandability of the property and its application to the HI teams for *Q1*), and finally, the mean and standard deviation of the scores each team obtained (*Q1*).

Each of the *Essentials* (the main conditions for a solid platform for human teams) was considered important for HI by *at least* 80% of the participants. Among the *Enablers* (team growth accelerators), the properties concerning a *Sound Structure* were considered important, but *Whole Task* scored lower (60%). Similarly, *Coaching* properties and *Rewards and Recognition* were on the 60% mark. Overall, all 22 properties were considered important for HI by at least 60% of the participants. Twelve properties (about half of all the properties) were considered important for HI by all the participants (100%).

> 👍 *Observation 1*
> Each property from the TDS was considered as important for HI teams by the majority of the participants.

Although all properties have some importance for HI teams, not all are directly applicable in the way expressed in TDS. We discern directly applicable properties by identifying those deemed both *important* (*Q3*) and *well understood* by *all* the participants. We treat these properties as *directly applicable* to HI teams, and highlight them in green in Table 2. Yellow cells indicate properties that were considered important by at least 80% of the participants but not directly applicable, and the red cells indicate properties that received a mixed feedback about their importance.

> 👍 *Observation 2*
> Properties concerning team structure, team composition, and team goals are well understood for HI teams.

> 👍 *Observation 3*
> Eight properties characterizing effective human teams are both important and well understood for HI teams: *bounded* and *interdependent* team, right *team size*, *diversity* among members, a *clear* and *consequential* purpose, *autonomy and judgment* of members, and *team norms*.

To understand whether the eight properties in Observation 3 are well-exhibited in the considered HI teams, we examine the scores the teams obtained for these properties. These scores are shown in Table 2 and graphically illustrated in Figure 2 via a spider chart. Notably, of the eight properties (underlined in Fig. 2), none of the HI teams reflected six properties very well, i.e., they did not receive an average score higher than 4. The two exceptions are *Clear (Compelling Direction)* for which only the Warehouse team reported an average score of 4.2, and *Team Size*, for which the Entertainment and Research teams received an average score ≥ 4.

The lack of clarity of direction stemmed from the AI's limited context awareness (e.g., "the weather might affect the type of movie to watch", for the Entertainment team) and opacity regarding AI "information utilization and optimization metrics (for both human and AI members)". For *Team Size*, participants reported that 1-to-1 teams (Entertainment and Research) "can hardly be larger". However, these teams scored low on *Diversity*. Participants noted that, despite its importance for HI teams, there are contexts where "diversity may not be needed" or is less suitable (e.g., "for physical work, or in 1-to-1 teams where the AI is optimized for one user").

An interesting case is that of the *Interdependent* property. Although it is considered both important and well understood for HI teams, participants noted that in all human-AI teams, "humans can

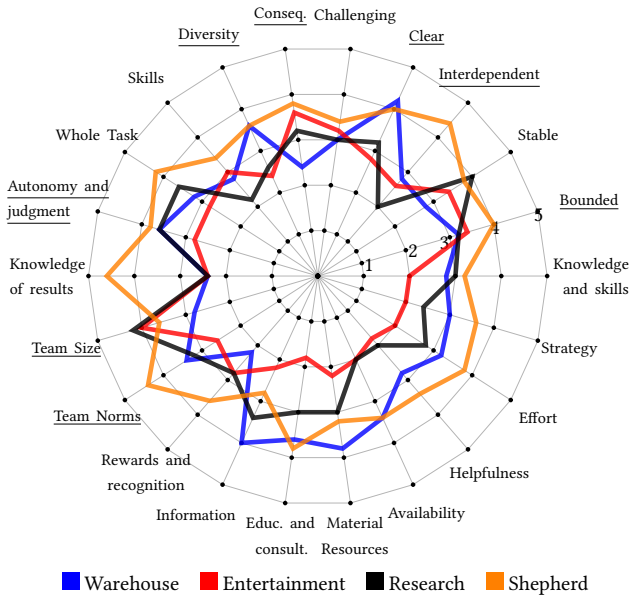

Figure 2: The mean scores on how well (*not at all* = 1, *extremely well* = 5) the HI teams reflect the TDS properties (*Q1*).

do their job without consulting the AI, while the AI cannot do any work on its own". Differently, the Shepherd team was considered to reflect the property more than very well, in line with known knowledge about human-dog interaction [43].

It is also important to understand properties that were not found to be directly applicable or well-understood for HI. Nine properties (yellow in Table 2) are not directly applicable to characterize HI teams, although they were perceived as important for HI by at least 80% of the participants. Additionally, five properties (red in Table 2) are neither directly applicable nor clearly important for HI. These properties received mixed feedback about their importance.

Notably, all the properties with mixed evaluation about importance were also associated with difficulties in their understanding or evaluating them in relation to HI teams. In the following, we discuss the difficulties that the participants' highlighted about these properties for HI teams, which lead us to the following observation.

Observation 4

Properties concerning social or (inter-)personal aspects (with the exception of *team norms*), and team processes, including self-development, and control over team strategies, are less understood for HI teams.

Participants noted that statements describing five properties (from all the TDS categories) made use of terminology that is either ambiguous or difficult to apply to non-human team members. This difficulty was raised, generally, with respect to the quantification of human-like abilities and behaviors that are not associated to current AI solutions. Specifically, the participants found it difficult to understand the following expressions when applied to the AI team members: *to fall into mindless routines* (*Strategy* property), *to share special knowledge* (*Knowledge and Skills*), *rewards and recognition* for AI members (*Rewards and recognition*), *stretching to accomplish goals* (*Challenging*), since "AI either can (within small error rate) or cannot do something", and *putting effort into the team* (*Effort*), since "non-humans lack intrinsic motivation toward team success".

Observation 5

The meaning of human-centric terminology poses difficulties when referring to non-human team members.

Participants also reported various multifaceted difficulties broadly related to the asymmetry between humans and AI.

First, participants raised the question as to whether teams possess a certain property if it is only exhibited by human members. This issue became apparent in properties such as *Strategy* and *Knowledge and Skills*, for instance, concerning lessons learned from experience. They also questioned whether algorithm updates should be considered as changes in team membership (*Stable*), "given that the [team] interactions may change dramatically".

Second, the context dependency of certain properties emerged as a recurring challenge. For instance, participants observed that in some scenarios, e.g., in the Warehouse team, AI replacements don't significantly impact (*Bounded*), whereas in others, e.g., Entertainment team, "shared accounts or simultaneous usage (e.g., group movie watching)" could alter team membership. Similarly, they noted that when dealing with HI systems, "trustworthy indicators are not necessarily only related to performance" and task execution (*Knowledge of results*), and that the importance of properties like *Autonomy and Judgment*, *Stability*, and *Team Size* and *Diversity* depends on the specific HI team and its objectives.

Observation 6

Properties characterizing human teams do not fully address the inherent human-AI asymmetry of HI teams.

Some participants found that the *Coaching*-related properties did not necessarily apply to HI teams. Others could not link coaching to anything existing in current systems, except FAQs, help-desks, or tutorials, which "still leave up to the human to learn how to use the AI better". Some participants noted that *Reward and recognition* are "possibly problematic, as the system can impose its task allocation and teamplay and not reward individual autonomy and creativity".

Observation 7

The relevance of *coaching*, and *reward and recognition* remains uncertain in the context of HI teams.

5.2 RQ2 (Effectiveness)

An important aspect of the TDS is measuring the effectiveness of human teams through seven measures (described in Section 3). Our study investigated their importance for HI teams (RQ2).

Figure 3 shows stacked bar charts that illustrate the responses of 14 participants to question *Q4* which addresses RQ2 (one of the 15 total participants did not answer this question).

For each measure, the most common response was *Very important*. Each measure was considered at least *Moderately important* for measuring the effectiveness of HI teams by at least 85% of the 14 participants. Only one participant evaluated user satisfaction (*E1*) and satisfaction of members with their relations (*E5*) as *Not at all important*. The appropriateness of performance strategies (*E2*) and the quality of group processes and team interactions (*E4*) were considered at least moderately important by all the participants.

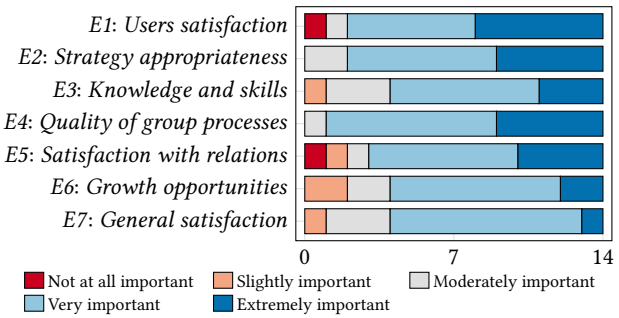

Figure 3: Effectiveness measures for HI teams (*Q4*).

Observation 8

Each measure of effectiveness of human teams from the TDS is important for HI teams.

We observe that among all the measures, those considered less important by some participants (e.g., *E5* and *E6*) explicitly refer to social and human-oriented facets. For example, *E5* concerns members satisfaction with their relations with other members. *E6* is described as the opportunity for team members to grow and learn over time. This is in line with Observations 4-6, pointing to challenges in comprehending social aspects and human-centric language for non-human members, and in addressing the human-AI asymmetry.

6 TOWARD A QUALITY MODEL

Based on the insights from Section 5, we lay the foundation for a quality model for HI teams. Our goal is not to develop a full-fledged quality model, but to formulate recommendations that can inform its future development. To facilitate this, we formulate our recommendations in terms of the two primary elements of quality models [31, 69, 77]: *quality attributes*, representing properties relevant for HI teams, and *quality measures*, quantifying quality attributes.

The recommended quality *attributes* summarize the insights from RQ1 (Section 5.1), and are based on the participants' feedback and the TDS statements (Table 1). The recommended *measures* address HI teams *effectiveness* (primary focus of the TDS) and follow from RQ2 (Section 5.2). In formulating our recommendations, we acknowledge the inherent human-AI asymmetry (Observation 6).

Figure 4 illustrates, in a similar manner to [31], our recommended (though not exhaustive) quality attributes and their refinements into lower-level attributes, and quality measures of effectiveness.

Observations 3 and 8 indicate that several properties of effective human teams were both considered important and well-understood by all the participants, and that effectiveness was considered important for HI teams. Based on this, we recommend the following.

Recommendation 1

Quality attributes of a HI team should include: *boundedness*, *interdependence*, *competency*, *purposefulness*, *initiative*, *normativity*, and *effectiveness*.

Boundedness, *interdependence*, *initiative*, *normativity*, and *effectiveness* follow directly from their corresponding properties from Observation 3. *Purposefulness* ties together the Clear and Consequential compelling direction properties from the TDS, and *Competency* ties together Team Size, Diversity and Skills.

Some participants highlighted the need for clarity of roles in HI teams. From this, we refine *Boundedness* into *Team structure transparency*—the degree to which team members have shared knowledge about team composition, roles and hierarchy. Further, from the Bounded property of TDS, we include *Members identifiability*—the degree to which team members can identify each other.

Following the TDS and the current literature [64], we refine *Interdependence* into *Mutual dependency*—the degree to which members depend on each other in achieving goals and tasks, and the degree to which they rely on *Communication* and *Coordination mechanisms*.

Participants recommended revising the Skills TDS property to explicitly mention "strengths and weaknesses of each human or AI, and whether the team solves deficiencies". From this, and from the importance of Team Size and Diversity, we formulate the *Competency* quality attribute, and refine it into *Skills comprehensiveness*—the degree to which the pool of skills in the team covers the needs of the team goals and tasks, and *Strengths and weaknesses transparency*—the degree to which team members have shared knowledge about each other strengths, weaknesses, and knowledge.

Purposefulness, in line with its corresponding properties from the TDS, is refined into *Objectives consequentiality*—the degree to which the team purpose is significant for the human members and stakeholders, and *Objectives transparency*—the degree to which team members have shared knowledge about team objectives and the link between team tasks and objectives—a need explicitly highlighted by some participants who mentioned the "knowledge asymmetry between AI and humans" in knowing when the team is doing well.

Initiative (from the Autonomy TDS property), is refined into *autonomy* and *proactivity*—the degrees to which the structure of the team enables members to operate independently and exhibit self-motivated behavior toward the accomplishment of team objectives.

In line with HI literature [1], some participants highlighted the importance of norm-aware AI members and of "well documented [and] acceptable behavior for AI". From this, *Normativity* (following from Team Norms), is refined into *Norm transparency* and *Norm awareness*, i.e., the degree to which team members have shared knowledge about team norms, and the degree to which members reason about team norms and adjust their behavior accordingly.

Effectiveness is refined, in line with the TDS, into *Task performance*, *Quality of group processes*, and *Members satisfaction*, respectively referring to the degree to which users of the team are satisfied with team's work, the degree to which the team becomes increasingly effective over time, and the degree to which the team contributes to the learning, growth and satisfaction of its members.

Our explicit investigation on effectiveness measures (RQ2, Section 5.2) led us to the following recommendation.

Recommendation 2

Effectiveness quality measures should include: *users satisfaction*, *strategy appropriateness*, *knowledge and skills*, *quality of group processes* and of *interactions*, *satisfaction with relations*, *growth opportunities*, *general satisfaction*.

We stress that none of these measures are exclusive to the AI component. This reinforces our recommendation that measures for HI teams should explicitly encompass the entire team and focus on the quality of interactions and dynamics among its members.

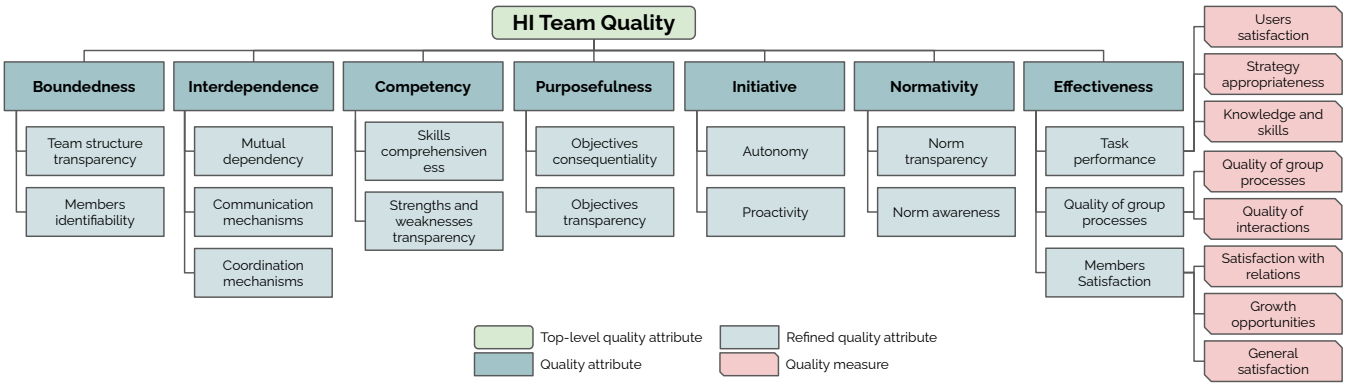

Figure 4: A summary of the recommended (though not exhaustive) quality attributes and measures for HI teams.

Our study also highlighted a number of difficulties in applying human-centric team characterizations to HI teams. Observations 5 and 6, for example, indicate that human-centric terminology and the asymmetry between humans and AI make it difficult to understand and apply some properties to non-human members. As a consequence, we recommend the identification of distinct quality attributes tailored to different types of HI team members to acknowledge the unique characteristics and goals of specific teams.

> *Recommendation 3*
> A quality model for HI teams should acknowledge and address the inherent human-AI asymmetry and the distinctive traits of each HI team.

Recommendation 3 suggests the need for different quality attributes based on team members and types. For example, in a warehouse scenario, *Members identifiability* may be pertinent for robotic but not for human team members, unless robots have unique characteristics. Likewise, in line with Observation 7, recognition and coaching might be relevant for human team members but not for non-human counterparts. The Strategy TDS property could lead to a quality attribute related to *adaptivity*, potentially only applicable to human team members due to the current AI systems limited flexibility in innovating work approaches. Finally, there's a chance to strengthen the link between the Information and Material resources TDS properties and Information Systems and training data, potentially closing the gap with existing AI systems quality models.

7 CONCLUSION AND DIRECTION

We proposed a team-oriented approach to HI, framing HI systems as HI teams. We investigated the adequacy and importance of human team properties and effectiveness measures to characterize HI teams. Results highlight the importance of several of these properties and effectiveness measures for HI teams. This led us to formulate recommendations for a quality model for HI teams, identifying seven high-level *quality attributes*, further refined into 16 specific ones, and eight *quality measures* for assessing HI team effectiveness.

By emphasizing a system-level, human-AI perspective, our preliminary quality model for HI teams provides an alternative perspective to existing technology-oriented taxonomies of HI and to existing AI quality models. Having such a quality model is not only important for assessing existing HI systems but also useful as a tool for designing new HI systems with quality in mind.

Our study highlighted the need for further research into addressing AI-human knowledge asymmetry, promoting hybrid teams interdependence, and integrating social or (inter-)personal factors in AI decision-making. Further, our study highlighted some limitations of the TDS when applied to hybrid teams. These included interpreting human-centric terminology for non-human team members, and the difficulty in addressing the inherent asymmetry between humans and AI. These aspects, not included in our preliminary model, require further study for incorporation into a quality model. Further work is needed to asses alternative team models (e.g., [73]). These can highlight additional relevant dimensions of HI teams, including aspects such as team cohesion [29, 54, 75], group engagement [58], and measures beyond effectiveness such as team process measures [35], privacy [40, 52], creativity [12, 53, 65] and dominance [32].

A limitation of our study is that it involved a limited number of subjects. We sought HI researchers as subjects since developing a quality model for an emerging concept like HI requires both the subject knowledge, and the willingness and time to engage in in-depth discussions. Finding such subjects is challenging. As evident from our experience, although we started with 50 researchers, eventually only 15 researchers completed the study. Yet, we synthesized interesting observations and recommendations, systematically. Importantly, the quality model we developed is a starting point, and it can be improved incrementally as new insights emerge.

Finally, we used the TDS as a trigger for discussing human team properties in relation to HI. Thus, the team scores in Table 2 are not a definitive assessment of the considered teams. The realization of a HI-Team diagnostic tool for quantitative assessment, and effective visualization strategies (e.g., Figure 2) for comparing HI teams over multiple dimensions remain interesting future directions.

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), and by the BOLD Cities initiative.

REFERENCES

[1] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen V. Hindriks, Holger H. Hoos, Hayley Hung, Catholijn M. Jonker, Christof Monz, Mark A. Neerincx, Frans A. Oliehoek, Henry Prakken, Stefan Schlobach, Linda C. 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 (2020), 18–28.

[2] 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.

[3] Rafael H. Bordini, Michael Fisher, and Maarten Sierhuis. 2009. Formal verification of human-robot teamwork. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, HRI 2009*. ACM, 267–268.

[4] Jeffrey M. Bradshaw, Virginia Dignum, Catholijn M. Jonker, and Maarten Sierhuis. 2012. Human-agent-robot teamwork. *IEEE Intelligent Systems* 27, 2 (2012), 8–13.

[5] Jeffrey M. Bradshaw, Paul J. Feltovich, Hyuckchul Jung, Shriniwas Kulkarni, William Taysom, and Andrzej Uszok. 2003. Dimensions of Adjustable Autonomy and Mixed-Initiative Interaction. In *Postproceedings of the 1st International Workshop on Agents and Computational Autonomy - Potential, Risks, Solutions, AUTONOMY 2003 (Lecture Notes in Computer Science, Vol. 2969)*. Springer, 17–39.

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

[7] Federico Cabitza, Andrea Campagner, and Luca Maria Sconfienza. 2021. Studying human-AI collaboration protocols: the case of the Kasparov's law in radiological double reading. *Health Information Science and Systems* 9, 1 (2021), 8.

[8] Mustafa Mert Çelikok, Frans A. Oliehoek, and Samuel Kaski. 2022. Best-Response Bayesian Reinforcement Learning with Bayes-adaptive POMDPs for Centaurs. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 235–243.

[9] European Commission, Content Directorate-General for Communications Networks, and Technology. 2019. *Ethics guidelines for trustworthy AI*. Publications Office.

[10] International Ergonomics Association. Congress and I.D. Brown. 1985. *Ergonomics International 85: Proceedings of the Ninth Congress of the International Ergonomics Association*. Taylor & Francis.

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

[12] Davide Dell'Anna and Anahita Jamshidnejad. 2022. Evolving Fuzzy Logic Systems for Creative Personalized Socially Assistive Robots. *Engineering Applications of Artificial Intelligence* 114 (2022), 105064.

[13] Davide Dell'Anna, Pradeep K. Murukannaiah, Bernd Dudzik, Davide Grossi, Catholijn M. Jonker, Catharine Oertel, and Pinar Yolum. 2024. *Toward a Quality Model for Hybrid Intelligence Teams - Supplementary Material*. https://doi.org/10.5281/zenodo.10593358

[14] Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, and Philipp Ebel. 2019. The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems. In *Proceedings of the 52nd Hawaii International Conference on System Sciences, HICSS 2019*. ScholarSpace, 1–10.

[15] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid Intelligence. *Business & Information Systems Engineering* 61, 5 (2019), 637–643.

[16] Paul M Fitts. 1951. *Human engineering for an effective air-navigation and traffic-control system*. National Research Council, Washington, DC.

[17] Gaku Fujii, Koichi Hamada, Fuyuki Ishikawa, Satoshi Masuda, Mineo Matsuya, Tomoyuki Myojin, Yasuharu Nishi, Hideto Ogawa, Takahiro Toku, Susumu Tokumoto, Kazunori Tsuchiya, and Yasuhiro Ujita. 2020. Guidelines for Quality Assurance of Machine Learning-Based Artificial Intelligence. *International Journal of Software Engineering and Knowledge Engineering* 30, 11&12 (2020), 1589–1606.

[18] Athina Georgara, Juan A. Rodríguez-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, AAMAS 2022*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 516–524.

[19] 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.

[20] Bahar Gezici and Ayça Kolukısa Tarhan. 2022. Systematic literature review on software quality for AI-based software. *Empirical Software Engineering* 27, 3 (2022), 66.

[21] Guido Governatori, Trevor J. M. Bench-Capon, Bart Verheij, Michal Araszkiewicz, Enrico Francesconi, and Matthias Grabmair. 2022. Thirty years of Artificial Intelligence and Law: the first decade. *Artificial Intelligence and Law* 30, 4 (2022), 481–519.

[22] Davide Grossi and Paolo Turrini. 2012. Dependence in games and dependence games. *Autonomous Agents and Multi-Agent Systems* 25 (2012), 284–312.

[23] Khan Mohammad Habibullah, Gregory Gay, and Jennifer Horkoff. 2023. Non-functional requirements for machine learning: Understanding current use and challenges among practitioners. *Requirements Engineering* 28, 2 (2023), 283–316.

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

[25] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. In *Proceedings of the 25th Pacific Asia Conference on Information Systems, PACIS 2021*. Association for Information Systems, 78.

[26] Thomas Herrmann. 2020. Socio-Technical Design of Hybrid Intelligence Systems - The Case of Predictive Maintenance. In *Proceedings of the First International Conference on Artificial Intelligence in HCI, AI-HCI 2020 (Lecture Notes in Computer Science, Vol. 12217)*. Springer, 298–309.

[27] 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.

[28] Sara Hoummady, Franck Péron, Dominique Grandjean, Delphine Cléro, Barbara Bernard, Emmanuelle Titeux, Loïc Desquilbet, and Caroline Gilbert. 2016. Relationships between personality of human–dog dyads and performances in working tasks. *Applied Animal Behaviour Science* 177 (2016), 42–51.

[29] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *IEEE Transactions on Multimedia* 12, 6 (2010), 563–575.

[30] 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.

[31] ISO/IEC 25010. 2011. ISO/IEC 25010:2011, Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models.

[32] Dinesh Babu Jayagopi, Hayley Hung, Chuohao Yeo, and Daniel Gatica-Perez. 2009. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 3 (2009), 501–513.

[33] 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.

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

[35] Anup K. Kalia, Norbou Buchler, Arwen H. DeCostanza, and Munindar P. Singh. 2017. Computing Team Process Measures From the Structure and Content of Broadcast Collaborative Communications. *IEEE Transactions on Computational Social Systems* 4, 2 (2017), 26–39.

[36] 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.

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

[38] 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.

[39] 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)*. Springer, 293–303.

[40] Abdurrahman Can Kurtan and Pinar Yolum. 2021. Assisting humans in privacy management: an agent-based approach. *Autonomous Agents and Multi-Agent Systems* 35, 1 (2021), 7.

[41] Hiroshi Kuwajima and Fuyuki Ishikawa. 2019. Adapting SQuaRE for Quality Assessment of Artificial Intelligence Systems. In *Proceedings of the IEEE International Symposium on Software Reliability Engineering Workshops, ISSRE Workshops 2019*. IEEE, 13–18.

[42] Chiyun Lee, Junxia Lin, Andrzej Prokop, Vancheswaran Gopalakrishnan, Richard N Hanna, Eliseo Papa, Adrian Freeman, Saleha Patel, Wen Yu, Monika Huhn, et al. 2022. StarGazer: A Hybrid Intelligence Platform for Drug Target Prioritization and Digital Drug Repositioning Using Streamlit. *Frontiers in Genetics* 13 (2022), 1–12.

[43] Chenyi Liao. 2021. *Applying the "human-dog interaction" metaphor in human-robot interaction: a co-design practice engaging healthy retired adults in China.*

Royal College of Art (United Kingdom).
[44] Joseph CR Licklider. 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics* 1 (1960), 4–11.
[45] Enrico Liscio, Michiel van der Meer, Luciano Cavalcante Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022. What values should an agent align with? *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022), 23.
[46] BF Lomov and VF Venda. 1978. Methodological principles of synthesis of hybrid intelligence systems. In *Proceedings of the International Conference on Cybernetics and Society, 1978*. IEEE, 1026–1030.
[47] 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.
[48] Antoine Marot, Adrian Kelly, Matija Naglic, Vincent Barbesant, Jochen Cremer, Alexandru Stefanov, and Jan Viebahn. 2022. Perspectives on future power system control centers for energy transition. *Journal of Modern Power Systems and Clean Energy* 10, 2 (2022), 328–344.
[49] Joseph Edward McGrath. 1984. *Groups: Interaction and performance*. Vol. 14. Prentice-Hall Englewood Cliffs, NJ.
[50] Joseph E McGrath, Holly Arrow, and Jennifer L Berdahl. 2000. The study of groups: Past, present, and future. *Personality and social psychology review* 4, 1 (2000), 95–105.
[51] Inge Molenaar. 2022. The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. *Computers & Education: Artificial Intelligence* 3 (2022), 100070.
[52] Francesca Mosca and Jose M. Such. 2021. ELVIRA: An Explainable Agent for Value and Utility-Driven Multiuser Privacy. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2021*. ACM, 916–924.
[53] Pradeep Kumar Murukannaiah, Nirav Ajmeri, and Munindar P. Singh. 2022. Enhancing Creativity as Innovation via Asynchronous Crowdwork. In *Proceedings of the 14th ACM Web Science Conference, WebSci 2022:*. ACM, 66–74.
[54] Marjolein C. Nanninga, Yanxia Zhang, Nale Lehmann-Willenbrock, Zoltán Szlávik, and Hayley Hung. 2017. Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017*. ACM, 206–215.
[55] Mark A. Neerincx, Jurriaan van Diggelen, and Leo van Breda. 2016. Interaction Design Patterns for Adaptive Human-Agent-Robot Teamwork in High-Risk Domains. In *Proceedings of the 13th International Conference on Engineering Psychology and Cognitive Ergonomics, EPCE 2016 (Lecture Notes in Computer Science, Vol. 9736)*. Springer, 211–220.
[56] An T. Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. 2018. Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, UIST 2018*. ACM, 189–199.
[57] Padmalata Nistala, Kesav Vithal Nori, and Raghu Reddy. 2019. Software quality models: a systematic mapping study. In *Proceedings of the International Conference on Software and System Processes, ICSSP 2019*. IEEE / ACM, 125–134.
[58] Catharine Oertel and Giampiero Salvi. 2013. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *Proceedings of the 2013 International Conference on Multimodal Interaction, ICMI 2013*. ACM, 99–106.
[59] Thomas A. O'Neill, Nathan J. McNeese, Amy Barron, and Beau G. Schelble. 2022. Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors* 64, 5 (2022), 904–938.
[60] Alison R. Panisson, Débora C. Engelmann, and Rafael H. Bordini. 2021. Engineering Explainable Agents: An Argumentation-Based Approach. In *Proceedings of the 9th International Workshop, EMAS 2021, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 13190)*. Springer, 273–291.
[61] Niccolo Pescetelli. 2021. A Brief Taxonomy of Hybrid Intelligence. *Forecasting* 3, 3 (2021), 633–643.
[62] 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.
[63] Aaron PJ Roberts, Leonie V Webster, Paul M Salmon, Rhona Flin, Eduardo Salas, Nancy J Cooke, Gemma JM Read, and Neville A Stanton. 2022. State of science: models and methods for understanding and enhancing teams and teamwork in complex sociotechnical systems. *Ergonomics* 65, 2 (2022), 161–187.
[64] Eduardo Salas, Terry L Dickinson, Sharolyn A Converse, and Scott I Tannenbaum. 1992. *Toward an understanding of team performance and training*. Ablex Publishing.
[65] R Keith Sawyer. 2010. Individual and group creativity. In *The Cambridge handbook of creativity*. Cambridge University Press, 366–380.
[66] Isabella Seeber, Eva A. C. Bittner, Robert O. Briggs, Triparna de Vreede, Gert-Jan de Vreede, Aaron Elkins, Ronald Maier, Alexander B. Merz, Sarah Oeste-Reiß, Nils L. Randrup, Gerhard Schwabe, and Matthias Söllner. 2020. Machines as teammates: A research agenda on AI in team collaboration. *Information & Management* 57, 2 (2020), 103174.
[67] Isabella Seeber, Lena Waizenegger, Stefan Seidel, Stefan Morana, Izak Benbasat, and Paul Benjamin Lowry. 2020. Collaborating with technology-based autonomous agents. *Internet Research* 30, 1 (2020), 1–18.
[68] Yash Raj Shrestha, Shiko M Ben-Menahem, and Georg Von Krogh. 2019. Organizational decision-making structures in the age of artificial intelligence. *California Management Review* 61, 4 (2019), 66–83.
[69] Julien Siebert, Lisa Joeckel, Jens Heidrich, Adam Trendowicz, Koji Nakamichi, Kyoko Ohashi, Isao Namba, Rieko Yamamoto, and Mikio Aoyama. 2022. Construction of a quality model for machine learning systems. *Software Quality Journal* 30, 2 (2022), 307–335.
[70] Dominik Siemon. 2022. Elaborating team roles for artificial intelligence-based teammates in human-AI collaboration. *Group Decision and Negotiation* 31, 5 (2022), 871–912.
[71] Roy Suddaby. 2010. Editor's comments: Construct clarity in theories of management and organization. *Academy of management Review* 35, 3 (2010), 346–357.
[72] Katia Sycara and Gita Sukthankar. 2006. Literature review of teamwork models. *Robotics Institute, Carnegie Mellon University* 31, 31 (2006), 1–31.
[73] Melissa A Valentine, Ingrid M Nembhard, and Amy C Edmondson. 2015. Measuring teamwork in health care settings: a review of survey instruments. *Medical care* 53, 4 (2015), e16–e30.
[74] Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. 2022. HyEnA: A Hybrid Method for Extracting Arguments from Opinions. In *Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence, HHAI 2022 (Frontiers in Artificial Intelligence and Applications, Vol. 354)*. IOS Press, 17–31.
[75] 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.
[76] Ruth Wageman, J Richard Hackman, and Erin Lehman. 2005. Team diagnostic survey: Development of an instrument. *The journal of applied behavioral science* 41, 4 (2005), 373–398.
[77] Stefan Wagner, Andreas Goeb, Lars Heinemann, Michael Kläs, Constanza Lampasona, Klaus Lochmann, Alois Mayr, Reinhold Plösch, Andreas Seidl, Jonathan Streit, and Adam Trendowicz. 2015. Operationalised product quality models and assessment: The Quamoco approach. *Information & Software Technology* 62 (2015), 101–123.
[78] 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.
[79] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to Complement Humans. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence, IJCAI 2020*. ijcai.org, 1526–1533.
[80] 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*. ACM, 114:1–114:28.
[81] Weishan Zhang, Huansheng Ning, Lu Liu, Qun Jin, and Vincenzo Piuri. 2021. Guest Editorial: Special Issue on Hybrid Human-Artificial Intelligence for Social Computing. *IEEE Transactions on Computational Social Systems* 8, 1 (2021), 118–121.
[82] Patrick Zschech, Jannis Walk, Kai Heinrich, Michael Vössing, and Niklas Kühl. 2021. A Picture is Worth a Collaboration: Accumulating Design Knowledge for Computer-Vision-based Hybrid Intelligence Systems. In *Proceedings of the 29th European Conference on Information Systems - Human Values Crisis in a Digitizing World, ECIS 2021*. Association for Information Systems, 1673.