Toward a Quality Model for Hybrid Intelligence Teams

Davide Dell’Anna
Utrecht University
Utrecht, The Netherlands
d.delannauuu.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
University of Amsterdam, NL
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.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:


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, 71]. 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, 61, 67], HI as a form of human-in-the-loop or AI-in-the-loop decision making [49, 69], HI as a type of collective intelligence [7], and HI as a design paradigm [8, 46, 75, 82].

Despite a wide array of interpretations, there is a lack of a quality model [78] that guides the systematic development and evaluation of HI systems [40]. Quality models (e.g., [32]) 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 [78].

Existing works on HI adopt a technology-centric perspective [14, 27, 62, 83], 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., [18, 21, 24, 70]) 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, 26, 68, 81]. 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 [58, 78].

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 empirically 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 [77], a well-known instrument employed in practice to diagnose the strengths and weaknesses of human teams.
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 [72] 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 [28, 35, 40] root back to the 50s [16, 45]. The term HI was first used in the late 70s in contexts of cybernetics [47] and ergonomics [10]. Nearly four decades later, the concept of HI has found applications in education [52], medicine [43], and computer vision [83].

Various taxonomies organize knowledge around HI and provide a framework for their design and evaluation [27, 49]. 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 [62] 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. [83] 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., [18, 21, 70]). Kuwajima et al. [42] integrate EU Guidelines for Trustworthy AI [9] into the ISO25010 [32] standard for system and software quality models. They propose new quality characteristics, such as controllability and explainability. Habibullah et al. [24] 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, 22, 80]. This motivates the human-agent team orientation of our work, complementing the existing literature.

There is extensive literature on teamwork (e.g., [3, 4, 56, 73]) and human-AI (including, human-robot and human-agent) teams [37, 60, 67]. Johnson et al. [34] 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. [81] investigate how complementary expertise between humans and agents play a role in creating team trust, and Kox et al. [39] develop strategies for agents to repair trust in human-AI teams. Wang et al. [79] design agents that provide explanations for their decisions to establish trust from humans in teams. Georgara et al. [19] develop explanation algorithms to clarify why certain teams can be formed with team formation algorithms and others cannot. Gervits et al. [20] make use of shared mental models to improve performance of human-robot teams. Paynadath et al. [63] 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 [64].

## 3 TEAM DIAGNOSTIC SURVEY

Human teams have been studied from many angles [23, 51]. Broadly, a team involves two or more members working interdependently toward a common goal [65] so that team activities result in collective success or failure, requiring accountability from all members.

IMO is a well-known conceptual model for teamwork [31, 50]. 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 [25]. Mediators include team processes [6, 48] and emergent states [11, 31, 38]. 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 [74] for a systematic review). The Team Diagnostic Survey (TDS) [77] is one such instrument. Table 1 provides an overview of the TDS. The TDS measures six conditions of effective teams’ design [77], 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 [77] characterizing each property. (R) stands for Reverse and indicates that the statement provides the opposite description of the property.

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

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 [74, 77] 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 discussion. 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, 57].

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 [29, 44].

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.
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 [77]. 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
We treat these properties as directly applicable 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%).

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).

<table>
<thead>
<tr>
<th>Property</th>
<th>Is important for HI (Q3)</th>
<th>Is well understood for HI teams (Q2)</th>
<th>Team’s scores (Q1)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Essentials</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real Team - Bounded</td>
<td>5 (100%)</td>
<td>Yes</td>
<td>3.2 ± 1.47</td>
</tr>
<tr>
<td>Real Team - Stable</td>
<td>4 (80%)</td>
<td>Yes</td>
<td>2.8 ± 0.75</td>
</tr>
<tr>
<td>Real Team - Interdependent</td>
<td>5 (100%)</td>
<td>Yes</td>
<td>2.8 ± 1.6</td>
</tr>
<tr>
<td>Compelling Direction - Clear</td>
<td>4 (80%)</td>
<td>No</td>
<td>3 ± 1.41</td>
</tr>
<tr>
<td>Compelling Direction - Challenging</td>
<td>5 (100%)</td>
<td>Yes</td>
<td>4.2 ± 0.75</td>
</tr>
<tr>
<td>Compelling Direction - Consequential</td>
<td>5 (100%)</td>
<td>Yes</td>
<td>2.4 ± 1.5</td>
</tr>
<tr>
<td>Right People - Diversity</td>
<td>5 (100%)</td>
<td>Yes</td>
<td>3.6 ± 1.36</td>
</tr>
<tr>
<td>Right People - Skills</td>
<td>4 (80%)</td>
<td>No</td>
<td>2.8 ± 1.17</td>
</tr>
<tr>
<td>Sound Structure - Whole Task</td>
<td>3 (60%)</td>
<td>No</td>
<td>3.2 ± 1.17</td>
</tr>
<tr>
<td>Sound Structure - Autonomy and judgment</td>
<td>5 (100%)</td>
<td>Yes</td>
<td>3.6 ± 1.36</td>
</tr>
<tr>
<td>Sound Structure - Knowledge of results</td>
<td>4 (80%)</td>
<td>No</td>
<td>2.4 ± 0.49</td>
</tr>
<tr>
<td>Sound Structure - Team Size</td>
<td>5 (100%)</td>
<td>Yes</td>
<td>2.8 ± 0.75</td>
</tr>
<tr>
<td>Sound Structure - Team Norms</td>
<td>5 (100%)</td>
<td>Yes</td>
<td>3.4 ± 0.8</td>
</tr>
<tr>
<td>Supportive Context - Rewards and recognition</td>
<td>3 (60%)</td>
<td>No</td>
<td>2.2 ± 1.17</td>
</tr>
<tr>
<td>Supportive Context - Information</td>
<td>5 (100%)</td>
<td>No</td>
<td>4 ± 0.63</td>
</tr>
<tr>
<td>Supportive Context - Education and consultation</td>
<td>4 (80%)</td>
<td>No</td>
<td>3.6 ± 0.8</td>
</tr>
<tr>
<td>Supportive Context - Material Resources</td>
<td>5 (100%)</td>
<td>No</td>
<td>3.8 ± 1.17</td>
</tr>
<tr>
<td>Coaching - Availability</td>
<td>3 (60%)</td>
<td>No</td>
<td>3.4 ± 0.49</td>
</tr>
<tr>
<td>Coaching - Helpfulness</td>
<td>3 (60%)</td>
<td>No</td>
<td>2.8 ± 1.17</td>
</tr>
<tr>
<td><strong>Enablers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Key Task</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effort</td>
<td>3 (60%)</td>
<td>No</td>
<td>3.2 ± 1.47</td>
</tr>
<tr>
<td>Strategy</td>
<td>5 (100%)</td>
<td>No</td>
<td>3 ± 1.1</td>
</tr>
<tr>
<td>Knowledge and skills</td>
<td>5 (100%)</td>
<td>No</td>
<td>2.8 ± 0.75</td>
</tr>
</tbody>
</table>

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

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
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 [44].

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.

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”.

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.

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”.

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.
Slightly (though not exhaustive) quality attributes and their refinements follow directly from their corresponding properties from the TDS, and sequentially, compelling direction properties from the TDS, and purposefulness.

Observation 3. Based on this, we recommend the following.

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 [32, 70, 78]: 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 the TDS property, addressing the inherent human-AI asymmetry (Observation 6).

Figure 4 illustrates, in a similar manner to [32], 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.

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 [65], 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.
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 assess alternative team models (e.g., [74]). These can highlight additional relevant dimensions of HI teams, including aspects such as team cohesion [30, 55, 76], group engagement [59], and measures beyond effectiveness such as team process measures [36], privacy [17, 41, 53], creativity [12, 54, 66] and dominance [33].

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), by the BOLD Cities initiative, and by the Health Holland epartners4all project.