Teamwork and Input Compositionality: UI Design for Taskwork or Teamwork?

ANONYMOUS AUTHOR(S)

The availability of AI-powered products and services has increased and become more accessible to people in their day-to-day lives. A significant portion of these products and services have come in the form of AI assistants, where the purpose of the AI system is to assist a human in a particular task such as driving or decision support. This type of human-AI teaming is one of the growing areas of research aiming at augmenting humans' capabilities and enhancing team performance; yet, the multifaceted nature of human-AI interaction poses challenges to researchers and practitioners. In particular, there are inconsistencies in the use of vocabularies to describe teamwork, including cooperation, collaboration, and coordination (3Cs), making it difficult to transfer research findings from one study to another across the literature. We propose a new approach to classifying teams based on input compositionality, enabling a more consistent use of the 3Cs. Our approach highlights that compositional control collaborative teamwork where inputs from team members are combined to produce one output, working intimately as a team is a unique, interesting type of team setting. To better understand this type of teamwork for human-AI teams, we conducted a game experiment where participants were asked to maneuver a spacecraft working in concert with an agent while simultaneously performing a secondary cognitive task. Experiment results underscore the importance of providing information about a compositional control collaborative task to humans working with a low capable agent. The results implied (albeit not conclusively) that conveying in a timely manner the degree to which an agent needs help from humans could improve team performance without a significant increase in humans' workload. The findings and lessons learned from the experiment are expected to be transferable to other compositional control collaborative teamwork settings, including AI-enabled automated driving.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**; **Empirical studies in HCI**.

Additional Key Words and Phrases: Human-AI Teamwork, input compositionality, collaborative interface design

ACM Reference Format:

Anonymous Author(s). 2018. Teamwork and Input Compositionality: UI Design for Taskwork or Teamwork?. In . ACM, New York, NY, USA, 29 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recent advances in artificial intelligence (AI) technologies, including Large Language Models (LLMs), have enabled new AI-powered products and services to become more accessible to people, allowing us to search for what we want to know quickly by asking questions (e.g., ChatGPT [1] [1]), to do tedious and repetitive tasks more efficiently (e.g., Github Copilot[2]), and/or to enjoy some creative work (e.g., DALL-E 3[3] is expected to offer even better experience). It has been several years since AlphaGo beat a professional human Go player [95], and now AI remarkably exhibits human-level performance in a strategy game requiring more sophisticated skills such as cooperation and negotiation

[1]https://chat.openai.com/ (Accessed on October 21, 2024)
[2]https://github.com/features/copilot (Accessed on October 21, 2024)
[3]https://openai.com/dall-e-3 (Accessed on October 21, 2024)

Manuscript submitted to ACM

[32]. Increasingly capable AI is also expected to interact with humans and play a more important role in high-stakes domains such as healthcare [81, 106], self-driving cars (e.g., Tesla[4], Waymo[5]), and cybersecurity [44, 53].

While these AI-assisted systems have prevailed in our daily lives, several concerns have arisen. One is centered on ethical issues in a wide range of sectors such as academia [20, 27], art [22, 31], social media [102], etc., warranting well-grounded regulations along with actionable practices [61]. Another pitfall lies in overreliance on AI while humans are interacting with it, and the importance of possessing appropriate reliance is more pronounced in more safety-critical scenarios and systems. For instance, although AI-assisted image diagnosis of cancer skin shows high accuracy, its performance is still hampered by multiple factors such as the quality of datasets [40]. Because of issues like these that reflect users' uncertainty in how to gauge trust in AI systems, advanced AI systems are being widely deployed first where they can augment human capabilities rather than replace them [23, 46, 50]. In applications such as decision support and self-driving, AI technologies augment human performance not by performing tasks that humans are incapable of, but by performing human tasks with increased accuracy, consistency, and endurance. Since the AI is performing tasks the human is also capable of, the trust dynamic in these types of applications can be different from other applications of AI technology because the AI performance is understandable and the human can generally take over control and perform the task themselves if they do not agree with the AI approach. Therefore, it is imperative to design interactions between humans and such agents to accomplish the ultimate goal.

Human-AI teaming is one of the key research areas in human-centered AI [11, 92] and precedents of human-AI interaction design such as the dynamic Situation Awareness-based Agent Transparency (SAT) model [13] and interdependence analysis [48] paved the way for enhancing humans' capability and elevating team performance. Yet, additional research work is necessary to further human-AI teaming research. For example, there is an inconsistency in the use of vocabularies to describe teamwork between humans and synthetic agents, including key terms such as collaboration, cooperation, and coordination (3Cs). A variety of teamwork settings are described differently using the 3Cs across the studies, making it difficult to transfer design insights from one research study to another due to this inconsistency. Thus, it is helpful to construct a frame of reference to discuss teamwork scenarios between humans and synthetic agents in a more consistent and coherent manner.

Another promising research avenue is to establish design principles and guidelines for improving human-AI team performance. The Human-Computer Interaction (HCI) community has been proposing design rules [93], principles, [73] and heuristics [71], helping the practitioners make more informed decisions on UI design and improve usability of systems, products, and/or services. Having such design principles and guidelines in the human-AI teaming community is expected to further advance research on improving teamwork between humans and AI.

This paper is intended to address these gaps. First, we propose a new approach to classifying teams based on how inputs to action are used and affect the system. We introduce two types of teams: one is a compositional control (inputs are combined) team, and the other is a non-compositional control (inputs are independent) team. Our new classification of teams clarifies distinctions between different types of teamwork activities and enables a more consistent use of the 3Cs. Also, our approach to mapping out different types of teams underscores that a compositional control is unique to collaborative teams, which makes it an interesting and relevant area of study (e.g., autonomous driving, collaborative decision-making). Therefore, our main focus in this paper is on compositional control collaborative human-agent teams.

Our study was originally inspired by a social physics study done by Pentland [78] that revealed the effects of consistent patterns of communication interactions between team members in effective human teams. We similarly

[4]https://www.tesla.com/autopilot (Accessed on October 21, 2024)
[5]https://waymo.com (Accessed on October 21, 2024)

aimed to identify effective teamwork patterns in a compositional control collaborative human-agent team setting. We conducted an experiment by employing a moon lander compositional control collaborative game, where participants were asked to collaborate with an agent on a moon lander maneuvering task while performing a secondary cognitive task concurrently. Through the experiment, we tested and compared the effectiveness of a taskwork-oriented UI and a teamwork-oriented UI in the compositional control collaborative team setting. More specifically, we hypothesized that the teamwork task of communicating with humans about when an agent needs help from them was key to achieving high performance in the task. We expected human-agent teams using such a UI design to produce high team performance while showing consistent interaction patterns. Although we did not confirm this hypothesis, the results indentified fruitful insights into patterns of interactions, relationships between agent capabilities and UI designs, and design considerations for teamwork-oriented UI.

In summary, this paper offers the following three contributions:

- We introduce a new approach to classifying teams based on how inputs from team members are used; compositional and non-compositional control teams
- We define cooperation, collaboration, and coordination in relation to task goals and input compositionality, highlighting a compositional control collaboration team as a unique, interesting type of team
- We report experiment results highlighting the importance of presenting information in a compositional control collaborative task when humans work with a less capable agent and hinting that conveying how much help an agent needs from humans in a timely fashion could help to amplify team performance without overloading humans' cognitive resources.

2 RELATED WORK

2.1 Teams and Teamwork

The literature [30, 34, 47, 51, 58, 87] offers definitions of teams whose commonalities include four key aspects: members/individuals, common/shared goals, interdependence, and roles or functions; yet, there is no consistent use of these terms [82]. Researchers have proposed different ways to classify teams using various factors, including skill, authority differentiation and experience of working as a team [45], the amount of communication and task interdependence [79], and perceived human-likeness, autonomy, and interdependence [62], process and risk [84], leadership assumption, plan coordination, and task allocation [99], to name a few.

When team members work together, two important factors are thought to contribute to team performance: taskwork and teamwork [29]. Taskwork refers to components that individuals independently perform, not necessarily requiring interdependent interactions between team members [86]. In contrast, teamwork has been described as interactions, mechanisms, or activities performed by individuals who work together toward goals. Within these definitions there are some differences: some include the notion of coordination [9], interdependence [47], or both [86]. Nissen et al. [72] tied taskwork and teamwork with cooperation and collaboration, respectively, meaning that collaboration requires interdependence between team members whereas individuals work on separate assignments in a more independent manner in a cooperation setting.

2.2 Cooperation, Collaboration, and Coordination (3Cs) in Teams

Cooperation, Collaboration, and Coordination (3Cs) are terms typically used to describe the behavior of teams. However, these terms tend to be used interchangeably as well as to be defined in different ways, introducing an inconsistency across

studies [12, 17, 59, 66]. In the literature on human teams (e.g., management), commonalities of collaboration appear to include: two or more members, working together, and a common goal [4, 24, 109] although some present unique perspectives (e.g., work together on tasks that cannot be accomplished individually [67]). Definitions of cooperation tend to contain the notion of benefit [6, 8, 37, 60] and voluntary action [8, 19]. As mentioned earlier, Nissen et al. [72] addressed distinctions between cooperation and collaboration associating the two terms with taskwork and teamwork, respectively.

When compared to collaboration and cooperation, definitions of coordination seem to be slightly more consistent across studies, where key notions include synchronizing the timing of actions and managing interdependencies [63, 64, 85]. Studies in the literature on human teams, including disaster management have examined the relationship between the 3Cs using different dimensions (e.g., following of common goals, shared resources, and shared risk [65], levels of information exchange and partner asymmetry [54], authority and resources & risks [100], and resources and information flow [39]).

Inconsistencies in use of the 3Cs is also found in the literature of teams with humans and synthetic entities [59, 66], and as seen in the literature on human teams, collaboration and cooperation are prone to be used interchangeably [7, 38, 52, 94, 101, 105]. Some Human-Robot Interaction (HRI) studies addressed a distinction between cooperation and collaboration; Bi et al. stated *"During cooperation tasks, robot and human partners interact without the need to know what the other is doing in a shared task. However, during collaboration tasks, both partners should communicate with and understand each other ..."* [7, p. 114]. Also, Kolbeinsson et al. distinguished them as follows: *"Cooperation, on the one hand, is described as a sequence of actions towards a shared goal, that each person is doing independently via subtasks towards the shared goal ... Collaboration, on the other hand, is described as a sequence of shared actions towards a shared goal"* [52, p. 453]. Sidji et al. [94] acknowledged the inconsistent use of cooperation and collaboration in the field of human-AI interaction as well. From their perspective, human-AI cooperation strives for improving joint welfare between humans and machines [21] whereas human-AI collaboration is under the umbrella of human-AI cooperation, where AI agents are expected to serve as assistants and help humans achieve their goals [94]. With regard to the relationship between drivers and highly automated vehicles, Lee et al. [59] also found inconsistency in the use of the 3Cs and addressed their relationships by accounting for resilience and time scale.

2.3 Patterns of High-Performing Teams

The literature on social physics offers fruitful insights into improving teamwork. Pentland [78] investigated characteristics of productive human teams employing wearable equipment to measure sociometric data, showing that effective human teams exhibit consistent patterns of communication between team members. Other studies leveraging sociometric data also found patterns of interactions exhibited by productive teams working on a wide array of tasks [104] and collective design [108]. Understanding such patterns of interactions helps to establish actionable solutions for amplifying teamwork (e.g., introducing a common coffee break room to increase interactions can improve the performance of a less effective team [78]).

In the literature on human-AI teams, there are studies suggesting characteristics of high-performing teams include the communication strategy [15], type [10, 42] and style [57], team coordination stability [25], and human participation in control [70]. Additionally, some research has applied social exchange theory aiming at dissecting interactions between humans and agents to improve team performance [15, 16]. It would be promising to understand differences in interactions between high- and low-performing teams, which could allow us to leverage patterns of interaction to make a more informed decision on UI design and evaluation to improve team performance.

2.4 UI Design Considerations for Human-AI Teaming

Researchers have examined how and what information is presented to humans to improve team performance in an AI-assisted decision-making task. Zhang et al. [107] reported that presenting an agent's confidence level helped with the human's trust calibration process in a human-AI decision-making task; still, they highlighted the importance of whether humans can bring and utilize their unique knowledge to correct the agent's errors. Bansal et al. [3] found that, with AI explanations, humans tended to rely on AI recommendations even when they were incorrect and stressed the importance of providing humans with informative explanations instead of just convincing ones. Prabhudesai et al. [80] observed increased time to make a decision when uncertainty information was presented to humans, resulting in a decrease in overreliance. Schemmer et al. [89] introduced the appropriateness of reliance by accounting for whether humans change their initial decision after receiving AI advice and reported the positive effect of explanations on human-AI decision-making performance.

Another rich body of research on human-AI teaming is cooperative driving, and studies have investigated types of information needed and presentation methods to accomplish safe driving. Although some studies reported that showing uncertainty information of automated system led to better human's response to automation failures [5, 43], others suggested a need for paying extra attention to the representation and presentation of uncertainty/confidence level information in automated driving. For example, Kunze et al. [56] tested an anthropomorphic representation to convey automation's uncertainty in automated driving and observed safer take-over performance at the cost of an increase in workload. Peintner et al. [77] also pointed out a potential pitfall when confidence levels are displayed with numeric values in automated driving because such a representation can overtax and/or unsettle humans when a low confidence level was presented. With a focus on SAE Level 4 automated driving [74], Peintner and his colleagues [76] explored different modalities to convey uncertainty information, revealing that participants preferred vibrotactile and auditory channels rather than a visual representation. Their study also found a trend where the participants wanted an option to collaborate[6] with and change automation's behavior even in Level 4 or 5 automated driving [74].

3 TEAM CLASSIFICATION BASED ON INPUT COMPOSITIONALITY

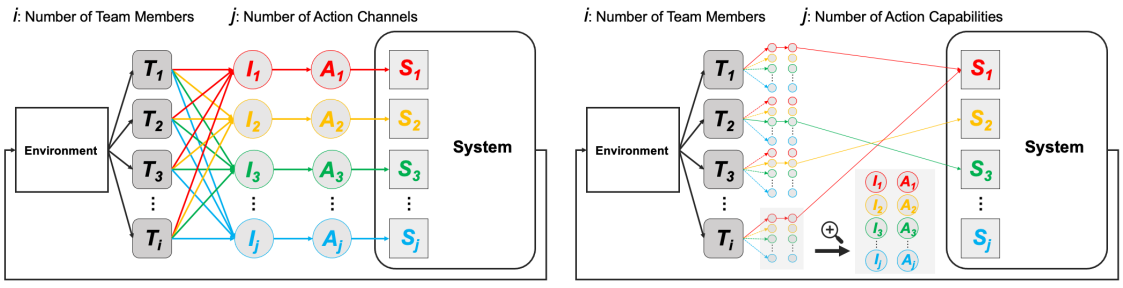
3.1 Compositional and Non-Compositional Control Teams

Our definition of a team is consisting of multiple entities who engage in activities interdependently to achieve goals while performing their roles in a dynamic, timely, and context-specific manner. In this paper, we use the term "agent" to refer to intelligent synthetic entities with the ability to sense and collect data, communicate with humans, take actions, and optionally to learn and evolve within their environment [18].

Momose et al. [69] dissected different forms of teams with a focus on how inputs from team members are used and affect the system. Here, we extend their concept and introduce two types of teams: **compositional control** (Fgure 1a) and **non-compositional control** (Fgure 1b) teams. The types are based on the concept of an action channel, which is the means by which actions are taken and produce an effect on the system. Team members in a compositional control team affect the system using the same action channels, and type/magnitude of the action is determined by the simultaneous composition of inputs from all team members. In contrast, in a non-compositional control team setting, team members can have different action channels that each affect the system independently.

Key distinctions between compositional and non-compositional control teams are threefold. First, adding team members can improve task efficiency of a non-compositional control team because each additional team member can

[6]Note that [76] refers to this a cooperation rather than collaboration



(a) Compositional Control Teams (b) Non-Compositional Control Teams

Fig. 1. (a) In a compositional control team, team members (T_i) simultaneously possess the same action channels where inputs from all members are composed. The team has one single input per action channel, generating one single action (A_j) affecting one of the states (S_j) in the system. (b) Team members in a non-compositional control team have the same action capabilities or unique action capabilities, and their actions affect system states independently.

affect the system independently. For example, adding one more member to a cleaning crew decreases the amount of time needed to clean, because the new member can clean simultaneously and independently with the rest of the team. In contrast, a compositional control team combines all teammates inputs in order to produce one action, and therefore increasing the number of team members may help to improve their input quality (e.g., a team comes up with a better business plan by consolidating inputs from all team members) although the efficiency is not significantly changed.

Second, a compositional control team can take the form of a non-compositional control team, but not vice versa. An input composition function in a compositional control team shares the same notion as the discretionary task in Steiner's task taxonomy [96], meaning that team members can determine how to combine inputs from each member through negotiations or based on a pre-defined mixture ratio. Therefore, one member may wish to possess a full control authority for one specific action, and likewise, others can also have a full responsibility for different actions channels. In such a scenario, the team behaves in a non-compositional team manner as each has different action capabilities. Still, the team can reconfigure the input composition functions at any time, and this flexibility and reversibility are not achievable in a non-compositional control team.

Third, hard interdependence, where goal achievement cannot occur unless another member accomplishes a subgoal, can occur only in a non-compositional control team setting. As some members may have unique action capabilities in a non-compositional control team, the team can face a situation where such members force the remaining team members to stand by until their tasks are complete. A compositional control system requires that team members are able to perform any task, thereby removing the need to wait.

3.2 Input Compositionality and 3Cs

With our approach to distinguishing types of teams based on the input compositionality, we propose definitions of the 3Cs. Figure 2 summarizes a relationship between the input compositionality and the 3Cs. When entities work together on a common goal, we define this situation as **collaboration**. Examples of compositional control collaboration include automated driving and human-AI decision-making (**Cell A** in Figure 2). As mentioned earlier, a team can determine how to blend inputs from all members in a discretionary task manner [96] and then generate one single team input per action channel. An example of non-compositional control collaboration (**Cell B** in Figure 2) is carrying and setting up

chairs in a conference room. All members have the same action capabilities and affect the system independently; a team working on an additive task [96] can fall into this category (e.g., the above mentioned cleaning crew example).

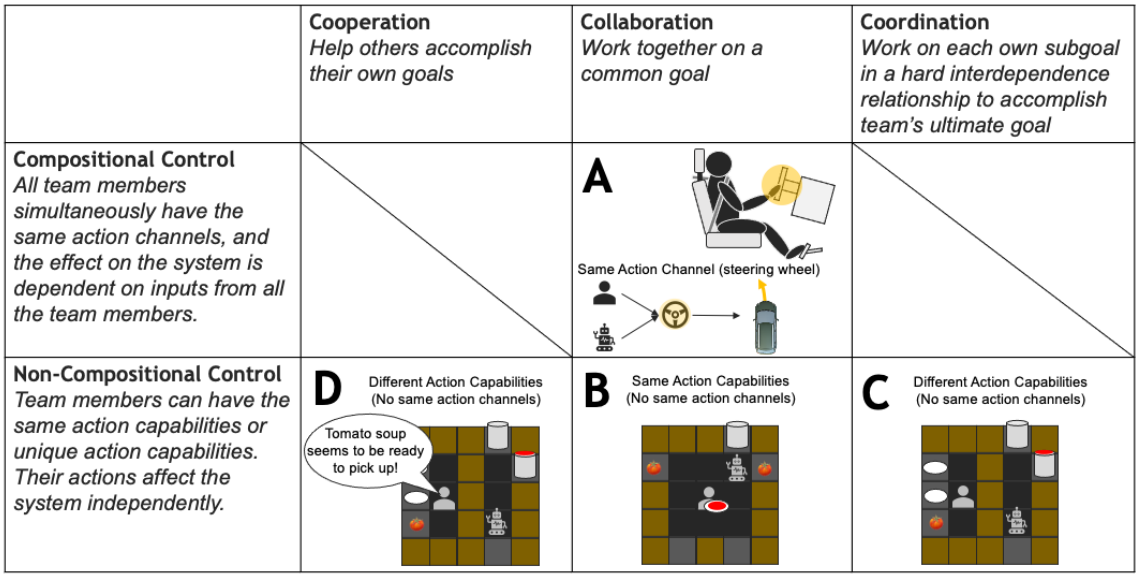


Fig. 2. Relationship between input compositionality and 3Cs. (A) Members in a compositional control team simultaneously have the same action channels (e.g., in self-driving, a steering wheel and gas pedals), and their inputs are composed, affecting the system (e.g., in self-driving, vehicle physics, including speed, attitude, etc.). (B) In a non-compositional control collaboration team, all team members have the same action capabilities (not the same action channels). Each contribution independently affects the system. (C) A situation where team members have unique action capabilities introduces a hard interdependence relationship [48], requiring coordination. (D) Cooperation is to be considered to be soft interdependence [48], improving team effectiveness and efficiency. Appendix A provides examples of compositional and non-compositional control teams with diagrams.

Let us use a cooking game [97] to explain the 3Cs in a non-compositional control team setting. In the cooking game, team members cook tomato soup and deliver tomato soup dishes to designated counters. Cooking tomato soup requires the following actions: picking tomatoes up, putting them in cooking pots, and delivering cooked tomato soup to the counters. Each entity can possess the same action capabilities. **Cell B** in Figure 2 shows a non-compositional control collaboration setting, where all have the same action capabilities and access to resources, but different action channels. When each member works on each entity's subgoal in a hard interdependence fashion to accomplish team's ultimate goal, we consider this situation to be **coordination** (**Cell C** in Figure 2). In **Cell C**, each member's action capabilities are limited due to the cooking room layout, inducing a hard interdependence (or forced coordination [97]) situation.

Cooperation is a situation where team members help others accomplish their own subgoals (**Cell D** in Figure 2), and soft interdependence comes into play. Soft interdependence acts as a catalyst to enhance team effectiveness and efficiency, meaning that while soft interdependence is not required, but optional, there is a potential for a team to work more effectively and efficiently by managing soft interdependencies [48]. In the example shown in **Cell D**, the entity in the left side does not have to let the other know about the fact that the tomato soup is ready to pick up. However, doing so may result in better team performance. Therefore, Observability, Predictability, and Directability (OPD) [48]

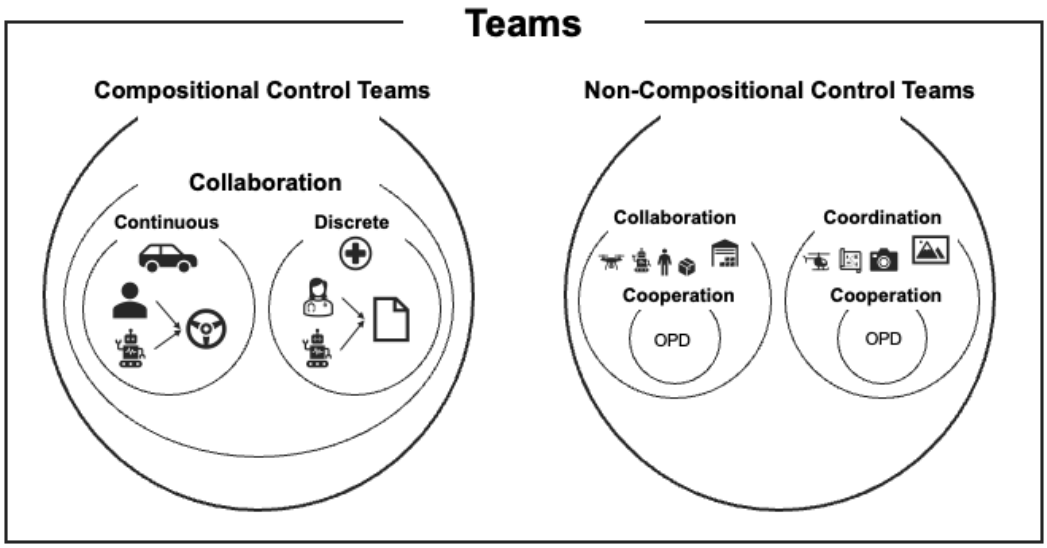


Fig. 3. Venn diagram of relationships between 3Cs in compositional and non-compositional control teams. In this figure, collaboration is defined as a situation where all team members have the same capabilities to fulfill post-conditions (i.e., change the post-condition state from false to true). Does cooperation occur only in coordination? What "help others accomplish their own goals" mean? Unique goals that can be accomplished only by them (ones have the unique capability to make the post-condition true)?

are expected to play a pivotal role in cooperation. It should be noted that soft interdependence also serves as a key element to improve team performance in compositional control collaboration as suggested by the automated driving studies focusing on conveying uncertainty information. (e.g., [56, 76, 77]); although an automated driving agent does not necessarily convey its uncertainty level to a driver, doing so is expected to help to achieve safer driving. However, we do not use cooperation for a compositional control team setting due to the fact that members in a compositional team always work together on a common goal, and each does not have each entity's subgoal.

Our approach is expected to allow the research community to discuss teamwork in different types of team activities in a more consistent manner. Furthermore, Figure 2 highlights that a compositional control collaboration team (**Cell A**) is a unique, interesting type of teams. Therefore, the rest of this paper presents our experiment investigating human-agent teamwork in a compositional control collaboration team setting.

3.3 Input Compositionality and Goal Structure

Another key distinction between compositional and non-compositional control teams is the existence of a hierarchical goal structure, which can be described using Planning Domain Definition Language (PDDL). Compositional control teams do not require pre- and post-conditions; in a compositional control team setting, all team members simultaneously have access to the same action channel, and the effect on the system is dependent on blended inputs from all the team members. Therefore, there are no incidents where one needs to wait until others complete their tasks and make their post-conditions true.

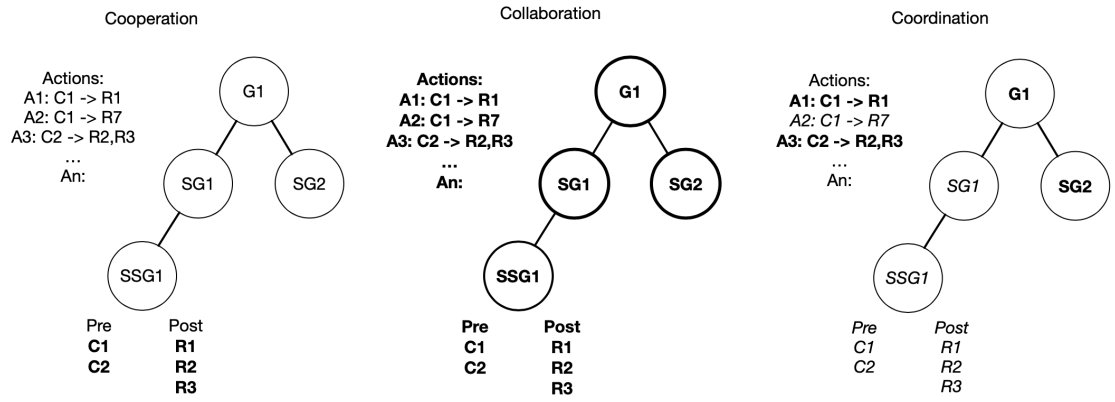


Fig. 4. Non-Comp Collaboration (all team members possess the same capabilities to fulfill post-conditions, non-comp coordination (at least one team member possesses a unique capability to fulfill post-condition(s), non-comp cooperation (can happen in both non-comp collaboration and coordination and helps others satisfy their pre-conditions)

In contrast, there is a goal hierarchy in a non-compositional control team setting, where satisfying one's post-condition serves as others' pre-condition, allowing them to proceed with their tasks. A key consideration is which post-conditions each team member can fulfill or make true. If all team members possess the same capabilities to fulfill all post-conditions, there is no hard interdependence because even only one entity can complete the team's ultimate goal (e.g., the rest of the team members are incapacitated) albeit less efficient. This is a unique type of non-compositional control collaboration team settings.

If one possesses a unique capability to fulfill a particular post-condition that cannot be made true by others, such a team encounters hard interdependence, and we refer this situation as coordination (non-compositional control collaboration-coordination).

From our perspective, cooperation is an act to help others fulfill their pre-condition(s) regardless of whether or not the helper possesses the same capability to make their corresponding post-condition(s) true (???). Figure presents an example, where each of the three entities can deliver and drop green, yellow, and red boxes respectively; each possesses its unique capability. If one team member delivering a green box finds one blue box on the way and comes across the teammate who can deliver blue one, the green box deliverer can inform the teammate of the blue box location, preventing the blue box deliverer from searching for other rooms (i.e., shortcut). In this case, the green deliverer satisfied the other's pre-condition (i.e., locates blue boxes). Therefore, cooperation is equivalent to soft interdependence (i.e., observability, predictability, directability), which is a nice-to-have feature and serves as a determinant of teamwork [48].

Our approach highlights that compositional control teams are a unique type team, and we argue that the very first type of human-ai teams that can be employed in real-world settings should be a comp-control collaboration team because of the input compositionality.

3.4 Role of Intent in Compositional Control Continuous Teams

In compositional control teams, all team members simultaneously have the same action channel(s), and their inputs are consolidated based on a pre-determined and/or dynamic composition function. Then, the team produces one single input for each action. Compositional control continuous collaboration teams are more likely to work in a more

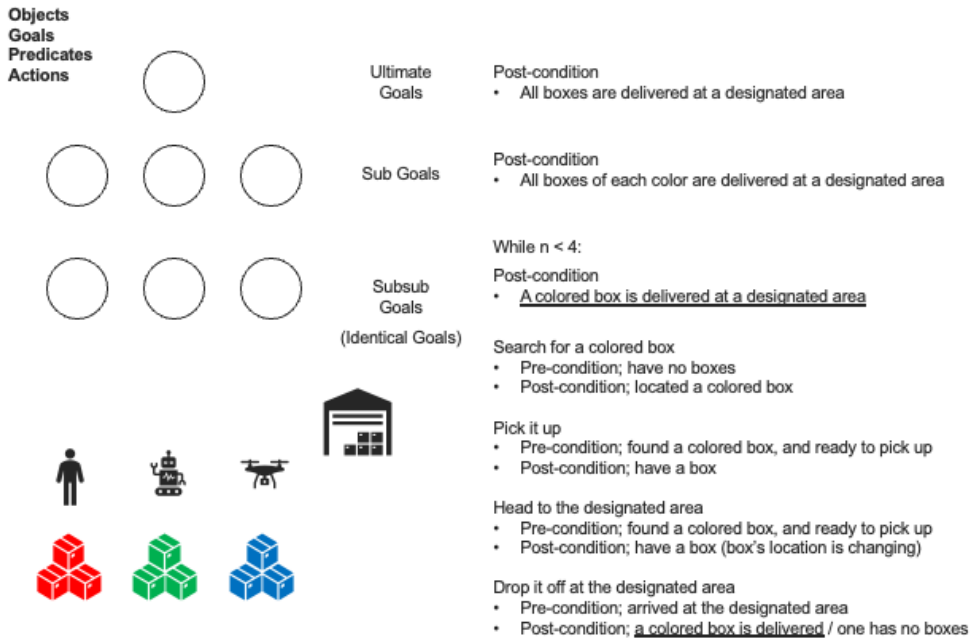


Fig. 5. Warehouse example to explain 3Cs in non-comp control teams

time-critical situation when compared to compositional control discrete collaboration teams, which highlights key design considerations for human-AI interaction.

In continuous collaboration, no inputs (i.e., absence of inputs from one or multiple entities) can still advance the feedback loop; physics/world environment does not wait for the team (e.g., the gravity force). In contrast, in the case of discrete collaboration, generally, once a team reaches to an agreement and generates an output, the team advances the next step. Given the these characteristics, providing explanation is less effective in compositional control continuous collaboration when compared to a discrete collaboration setting because while one explains something to others, the current state immediately becomes the previous step/state.

Instead, intent-oriented communication plays a significant role in improving compositional control continuous collaboration teamwork. Intent serves as a frame of reference, and the unique work structure of compositional control continuous collaboration allows team members to constantly observe others' behaviors and measure deviations between their original intent and what they are actually doing. Conveying intent eliminates the need for team members to predict others' behaviors, helping to reduce team members' cognitive workload. In the case where one observes large deviations from others' original intent, he or she can provide his or her inputs to rectify the current state.

It should be noted that in non-compositional control teams, ones can still indicate their intent, but others may not be able to observe entire process/action, preventing them from observing their behaviors and measuring deviations between their original intent and actual behaviors over the course of task.

- If some team members in a non-compositional control team do not have identical goals (i.e., they have unique goal(s)), such a team automatically falls into coordination as hard interdependency is introduced (i.e., you need to wait for others to make post-condition(s) of their unique goals true)

- Non-compositional control collaboration teams can have identical goals whereas compositional control collaboration teams always have the common (same?) goal.
- define different types of goals, including ultimate goal, common goals, identical goals, unique goals

4 IMPROVING COMPOSITIONAL CONTROL COLLABORATION TEAMWORK

Our study was initially inspired by one of the social physics studies done by Pentland [78], and we aimed to gain insights into what kinds of characteristics effective human-agent teams exhibit in a compositional control team setting. Momose et al. [70] identified some patterns of effective teams in a compositional control type human-agent team setting, although this was early-stage research and the results were inconclusive. Momose and his colleagues [69] extended their study by focusing on UI design for a compositional control collaborative task, finding that instant feedback on how human inputs contribute to task-based goal attainment is critical to a compositional control collaborative task. In this paper, we attempted to extend their work into UI design for a compositional control collaborative task by investigating the differences between UIs supporting taskwork (clarifying status and progress of the task) and UIs supporting teamwork (clarifying status and trends in teammate performance).

In the presented experiment, we investigated the following Research Questions (RQs):

RQ1 **Patterns of Effective Teams**: Do patterns of interactions inform how well teams collaborate?
RQ2a **UI Design**: Do UI designs change patterns of interactions between human and agent?
RQ2b **UI Design**: Does conveying agent intent help to improve compositional control human-agent teamwork?

5 EXPERIMENT

5.1 Overview

To answer our RQs, we conducted a 2 (agent capabilities; between-subject) × 3 (UI designs; between-subject) × 2 (secondary task difficulty levels; within-subject) experiment employing a collaborative human-agent game setting. The university's IRB reviewed and approved the experiment (IRB Number: 22-114). We preregistered the experiment protocol, including our hypotheses and data analysis method (AsPredicted #143062[7]).

5.2 Participants

We aimed at a target sample size of 192 suggested by G*Power [33][8]. Those who had a desktop/laptop with keyboard and a stable internet connection were eligible to participate in the experiment. To recruit participants, we employed convenience sampling, the university's general email forum, and announcements in classes. Some students received extra credit in their class upon the completion of the study.

5.3 Experiment Setting

Figure 6 shows a moon landing task game screen, consisting of two components: a lander maneuvering task (Figure 6a) and a secondary cognitive task (Figure 6b). The maneuvering task was a compositional control collaborative task, meaning that a human player and an agent shared the two action channels: a thruster control to change the lander's speed and a rotation control to change the lander's attitude. Both the human player and the agent were granted the same control authority level, meaning that either of them provided thruster and/or rotational input, the lander reacted with a

[7]https://aspredicted.org/1BQ_19H
[8]Effect size of 0.263 informed by our pilot study, alpha at 0.05, power of 0.80, number of groups of 8, and number of measures of 2.

full-throttle[9]. If one rotated the lander in a clockwise direction while the other provided an opposite direction input, their inputs offset each other, resulting in no rotational inputs. We employed the n-back task [75] as the secondary cognitive task, which was done by only the human player. The player was asked to respond to a target stimulus (see 5.4.3), and the n-back task was introduced to divert the player's attention from the maneuvering task. The goal of the moon lander game was to successfully land the moon lander in concert with the agent while simultaneously performing the n-back task. Both the maneuvering task and the n-back task contributed to the overall game scores, which were computed as follows:

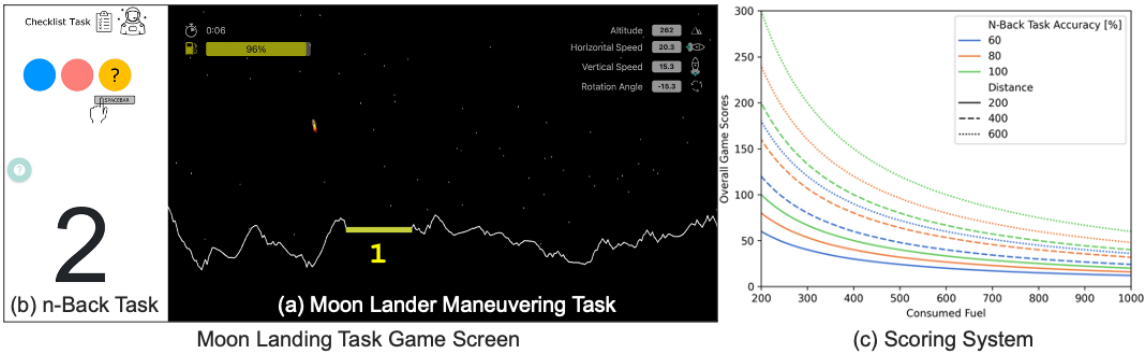


Fig. 6. The moon lander maneuvering task was presented as a compositional control collaboration task whereas the n-back task was simultaneously performed only by the human players. (a) Both the human player and the agent had the same action channels. The participants provided their thruster and/or rotational inputs using the up- and left/right arrow keys, respectively. The goal of the team was to safely land the lander on a landing pad. (b) The n-back task was employed to distract the human player's attention from the compositional control collaboration task; the n-back task was introduced as a checklist task to the participants. The human player was asked to respond in a timely manner whenever a target stimulus was presented. (c) Both the moon lander maneuvering task and the n-back task contributed to overall game scores. The participants were instructed to maximize them by successfully and efficiently performing the maneuvering task while accurately conducting the n-back task.

$$Overall\ Scores = \frac{Distance}{Consumed\ Fuel} \times N\text{-}Back\ Task\ Accuracy \tag{1}$$

where the *Distance* was measured from the initial lander position to the center of the landing pad and was randomized in every trial. The *Consumed Fuel* was the percentage of the amount of fuel used during the landing. The fuel was consumed when the human player and/or the agent provided thruster and/or rotational inputs; a greater amount of fuel was consumed when the team provided thruster inputs when compared to rotational inputs. If either the player or the agent provided a thruster input or a rotational input, the provided input was executed with a full throttle. Then, the *N-Back Task Accuracy* was calculated as follows:

$$N\text{-}Back\ Task\ Accuracy = \frac{H + CR}{H + M + FA + CR} \tag{2}$$

where H, CR, M, and FA are Hits (responded to a target stimulus in a timely manner), Correction Rejections (provided no inputs when a non-target stimulus was presented), Misses (failed to respond to a target stimulus in a timely manner), and False Alarms (responded to a non-target stimulus), respectively. The participants were instructed to maximize their

[9]The study by Momose et al. [70] employing a compositional control collaboration type task reported a trend where study participants did not change a control authority level between a human and an agent even with a capability to switch the control authority level in a real-time fashion.

overall game scores and were able to check their performance after each trial. Figure 6c illustrates the relationship between the three metrics in Equation 1.

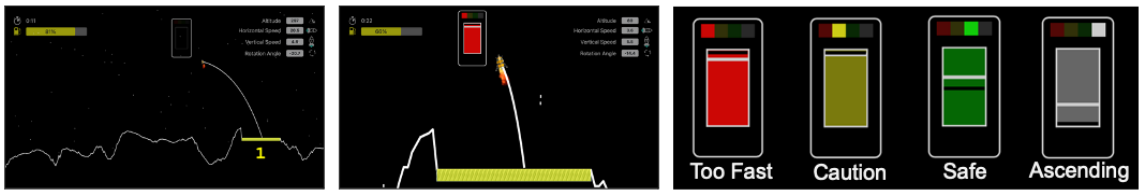
5.4 Independent variables (IVs)

5.4.1 Agent Capability (IV1). We prepared more- and less-capable agents showing different performance levels of the maneuvering task. Both agents were designed using a benchmark heuristic agent exhibiting 100% success rates. For the implementation of the more- and less-capable agents, we employed a notion of malfunction, meaning that we degraded the baseline heuristic agent by introducing a probability of taking no actions. We selected 79% and 89% chance of taking no actions for the more- and less-capable agents respectively, yielding an average agent-only landing success rate of roughly 90% and 40%.

5.4.2 UI Design (IV2). We tested three UI designs in the experiment: (i) only HUD information (hereafter referred to as baseline), (ii) lander physics, and (iii) agent state UI designs. The baseline design showed only flight-related text information on the game screen (see Figure 6a); the HUD information was presented across all the UI designs. The lander physics UI design (Figure 7) offered two features: a trajectory projection and a final approach aid. The trajectory projection was designed to display a free-fall trajectory (i.e., a white arc) by accounting for the lander's physics, allowing a player to predict where the lander impacts the surface with no additional thruster inputs (Figure 7a). If a player and/or an agent provided thruster inputs, the white arc responded to them accordingly. The final approach aid offered a thruster input recommendation alongside the trajectory projection during the final approach phase (Figure 7b). At the beginning of the landing task, the final approach aid was grayed out, indicating that it was currently disengaged. Once the lander approached the surface, a camera view zoomed in, and the final approach aid became activated. The final approach aid visualized the lander's descending speed using a horizontal bar with colors (Figure 7c). The red color indicated a too fast descending speed, and therefore the team needed to provide thruster input to reduce the speed. The yellow and green colors indicated that the current descending speed was within a safe range although the yellow state recommended some thruster inputs to reduce the speed. If the horizontal bar was in the gray color zone, the lander was ascending; the team needed to release the thruster inputs. The lander physics UI was considered to be a taskwork-oriented UI, meaning that it presented no information about agent's state, including its intent, but was designed to convey information about the landing task based purely on the current lander physics.

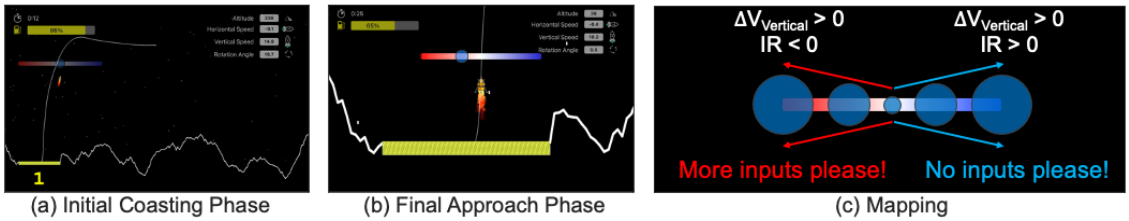
The agent state UI (Figure 8) was designed as a teamwork-oriented UI to convey information about the agent's internal state by generating the agent's intent path and introducing the notion of the agent's nervousness. At the beginning of a landing task, the agent computes the intended path to the landing pad (Figure 8a), which was generated based upon the benchmark heuristic agent exhibiting a 100% success rate. The agent tries to follow the generated path, and the agent and player are able to observe how well it maneuvered or how much it deviated from its intended path. As we introduced the notion of malfunction, there are deviations based on the capabilities of the agent.

The agent state UI also offered the agent nervousness visualizer in the vicinity of the lander (Figure 8b), which was intended to convey two pieces of information. First, the player was informed of whether the agent felt confident or nervous in performing a final approach maneuvering. Second, the agent nervousness visualizer let the player know whether the agent needed the player's help or not. The agent nervousness visualizer became active once the lander entered a final approach phase, where the camera view zoomed in. The size of a blue circle and its position were determined based on two metrics: $\Delta V_{\text{Vertical}}$ and Input Ratio (IR).



(a) Initial Coasting Phase (b) Final Approach Phase (c) Color Coding

Fig. 7. The lander physics UI was intended to serve as a taskwork-oriented UI. (a) In the first phase of flight, a white arc was displayed, representing a free-fall trajectory (i.e., if no thruster inputs were provided, the lander followed the free-fall trajectory). Once thruster inputs were provided, the white arc responded based on the lander physics. A final approach aid was greyed out and displayed in the vicinity of the lander in the first phase of flight. (b) Once the lander entered a final approach phase, the game screen zoomed in. Then, a final approach aid became activated. A white horizontal bar went up and down based on the current lander vertical speed. (c) The team was asked to land on a landing pad with the display in either the yellow or green state. The red state indicated too fast descending speed whereas the grey state was displayed when the lander ascended. The white horizontal bar had to be within the yellow or green state.



(a) Initial Coasting Phase (b) Final Approach Phase (c) Mapping

Fig. 8. The agent state UI was implemented as a teamwork-oriented UI design. (a) In the first phase of flight, a white arc was displayed, representing an agent's intended flight path computed based on the benchmark heuristic agent. The white arc was static, meaning that any inputs did not affect the agent's intended flight path. An agent nervousness visualizer was greyed out and displayed in the vicinity of the lander in the first phase of flight. (b) Once the lander entered a final approach phase, the game screen zoomed in. Then, the agent nervousness visualizer became activated. The size of the blue circle represented the agent's nervousness, and the position of the blue circle let the player know about the level to which the agent needed inputs from the player. (c) If the agent could not control the lander to match its intentions, the blue circle moved toward the left side while becoming larger, prompting the player to participate more actively in the maneuvering task. In contrast, if the lander deviated from agent intentions due to an excessive amount of human inputs, the agent asked the player to disengage by shifting the blue circle toward the right side and making the diameter larger.

The $\Delta V_{\text{Vertical}}$ is a difference between the current lander vertical speed and the agent intended vertical speed at the current altitude; the latter was determined based upon the benchmark agent. To compute the *IR*, we considered a history of key inputs from both entities within the last 99 steps using the following relationship:

$$Input\ Ratio\ (IR) = log_{10}\left(\frac{Human\ Inputs + 1}{Agent\ Inputs + 1}\right) \tag{3}$$

where *Human Inputs* and *Agent Inputs* were the number of inputs provided by the human player and the agent, respectively. Using Equation 3, we obtained a negative value when there were less human inputs, and more dominant inputs were provided from the agent, and vice versa. Then, we introduced the notion of urgency and computed it as follows:

$$Urgency = \Delta\ V_{Vertical} \times IR \tag{4}$$

With the urgency value, the agent nervousness visualizer let the player know if actions are required to return to the planned state (Figure 8c). If a large speed difference was observed, and *IR* took a negative value, the circle was becoming larger and moving toward the left (red) side, prompting the player to more actively participate in the maneuvering task. As the agent was designed to take no actions occasionally (i.e., the malfunction), it was prone to fail to reduce the vertical speed, and therefore the red side indicated a need for inputs from the player to compensate for the lack of agent's inputs. In contrast, the player saw a situation where the circle was getting larger and moving toward the right (blue) side. The blue side showed a trend where the lander was ascending due to unnecessary thruster inputs from the player (i.e., a positive value of *IR* induced a large $\Delta V_{\text{Vertical}}$), and therefore the player needed to release thruster inputs if a large circle was placed on the edge of the right edge.

The participants were instructed to make the blue circle small and attempt to keep it in the middle white range to achieve a successful landing. The agent nervousness visualizer was expected to enable the player to understand if inputs were required and how their inputs helped the agent reduce its nervousness level in a real-time manner.

5.4.3 N-Back Task (IV3). We employed 0- and 2-back tasks [75] as easy and more challenging secondary cognitive tasks respectively, serving as the within-subject factor. For the 0-back task, the player was told a target digit in advance, and once a trial began, a series of digits (from 0 to 9) were randomly presented on the left side of the game screen; a stimulus was displayed for 0.5 seconds, and there was a 2.25-second interval until the next digit appeared. The player was asked to hit a space bar if the current digit matched the prespecified digit. In the 2-back task, a target digit was not specified before a trial, and instead, the player was asked to provide a space bar input if the current digit matched the one displayed two steps previously.

5.5 Dependent variables (DVs)

5.5.1 Human Key Input Profile [-] (DV1). we kept track of a key input ratio between human and agent during a landing attempt. To compute the human key input ratio, we initially normalized the length of each trial. Then, we obtained the human key input ratio by dividing the number of human keystrokes (i.e., thruster and rotation control inputs) by the total number of keystrokes from the player and the agent per one normalized time step. Profiles of the human key input ratio in a trial were expected to illustrate how the human player participated in the maneuvering task. We clustered the profiles of human key input ratio to investigate differences in patterns of interactions between high- and low-performing teams.

5.5.2 Average Game Scores [-] (DV2). for each trial, we recorded the trial scores using Equation (1) and then computed averaged trial scores for each block of 10 trials. We recorded zero as trial scores for a failed trial.

5.5.3 Teamwork Workload [-] (DV3). we employed a modified version of the Team Workload Questionnaire (TWLQ) [90], which was administered in teamwork studies such as a UAV control task [91] and a collaborative decision-making task [41]. The modified version of the TWLQ consisted of Taskwork, Teamwork, and Team-Task Balancing components (Table 1). The taskwork and teamwork workload components asked the participants about their perceived workload level for the n-back task and moon lander maneuvering task, respectively. We computed the mean ratings of question items of the Teamwork component, ranging from 0 to 100.

5.5.4 Team-Task Balancing [-] (DV4). we also computed the mean of question items of the Team-Task Balancing component, ranging from 0 to 100.

Table 1. The modified version of the Team Workload Questionnaire (TWLQ) administered in the experiment. The TWLQ consisted of the taskwork, teamwork, and team-task balancing workload components. The taskwork and teamwork workload components asked the participants about their perceived workload level for the n-back task and moon lander maneuvering task, respectively In the experiment, the n-back task was introduced as a checklist task to the participants. The team-task balancing workload component was intended to capture the participants' overall experience of the moon landing task.

Component	Question
Taskwork (T)	T1: How mentally demanding was the checklist task? (not demanding at all - very demanding)
	T2: How hurried or rushed was the pace of the n-back task? (not hurried at all - very hurried)
	T3: How hard did you have to work to accomplish your level of performance? (not hard at all - very hard)
	T4: How irritated, stressed, and frustrated did you feel during the checklist task? (not at all - a lot)
Teamwork (TW)	TW1: How mentally demanding was the maneuvering task? (not demanding at all - very demanding)
	TW2: How much input was required from you to successfully land? (none - a lot)
	TW3: How difficult was it to understand the agent's maneuvering abilities? (Very easy - Very difficult)
Team-Task Balancing (TTB)	TTB1: How mentally demanding was conducting the maneuvering task and checklist task simultaneously? (not demanding at all - very demanding)
	TTB2: How difficult was it to determine which task to focus on to accomplish both the maneuvering and checklist tasks? (very easy - very Difficult)
	TTB3: How much did the agent help you determine which task to focus on to accomplish both the maneuvering and checklist tasks? (a lot - none)

5.6 Hypotheses

We established the following four hypotheses.

- H1 **Patterns of Effective Teams**: There will be differences in the overall team performance between groups clustered based on profiles of human key input ratio
- H2a **UI Design**: There will be more trials with the agent state UI in a high-performing team cluster than baseline and lander physics UI designs
- H2b **UI Design**: The agent state UI will enable higher game scores than baseline and lander physics UI designs
- H2c **UI Design**: Participants with the agent state UI will report a lower level of teamwork and team-task balancing workload

5.7 Experiment Procedure

The experiment was conducted remotely using the browser-based moon landing game. The length of the experiment was approximately 45 minutes. First, the participants signed an informed consent form and filled out a demographic questionnaire. Then, we randomly assigned the participants to one of the six groups (i.e., two agent capabilities for the three UI designs). Next, a familiarization session was presented, where the participants watched short tutorial videos and then performed familiarization trials of the moon lander maneuvering task and the n-back task first individually and then concurrently, allowing the participants to get acquainted with the tasks and the assigned UI design. During the familiarization trials, a familiarization agent that was 100% accurate was used to with all participants. Then, two

blocks with 10 trials each of the moon landing task were presented, where the participants worked on the moon maneuvering task with the assigned more- or less-capable agent using the assigned UI design while also performing the 0-back or 2-back task. We randomized the order of the n-back task difficulty levels for each block. After each block, the participants were asked to fill out the modified version of the TWLQ to report their perceived workload. The two blocks were followed by a debriefing session, where the participants were asked to provide their feedback on the assigned UI design via an open-ended question form. After submitting the form, the participants signed out from the online experiment.

5.8 Data Analysis

We used R (version 4.0.2) [98] for our statistical analysis and set the α level at 0.05. We first conducted a manipulation check for our n-back task treatment by running a t-test using the taskwork workload scores. If we would not confirm the effect of n-back task treatment via the manipulation check, we would consider the two n-back task levels to be identical. For H1 and H2a, we employed k-shape clustering using the dtwclust package [88] in R to generate four groups based on profiles of human key input ratio in successful trials across the four levels (i.e., 2 agent capabilities × 2 n-back task difficulty levels). After clustering, we ran one-way ANOVAs across the four levels to confirm H1, and if there were significant differences in game scores across the clusters, we performed follow-up analyses with the Bonferroni correction. For testing H2a, we carried out chi-square tests to investigate if there are significant differences in frequencies between UI designs in each cluster. To test H2b and H2c, we performed a three-way MANOVA using the MANOVA.RM package [36] for the overall game scores, teamwork, and team-task balancing workload scores. If appropriate, we conducted ANOVAs and pairwise comparisons with the Bonferroni correction using ARTool package [28, 103].

6 RESULTS

A total of 177 participants completed the entire experiment[10]. There were four participants whose trial data in one block were not recorded at all due to technical issues, and therefore we could not include them in our data analysis. Also, one participant exhibited no successful landing during the entire experiment, and therefore we also excluded the participant as we pre-registered. As a result, data from 172 participants aged from 16 to 59 years old (M = 22.0, SD = 5.74) were included in our data analysis.

6.1 Hypothesis Testing

Figure 9 shows (a) overall game scores, (b) taskwork, (c) teamwork, and (d) team-task balancing workload scores across all the conditions. With the taskwork workload scores, we carried out a Wilcoxon signed-rank test as the manipulation check to examine the effect of the n-back task, revealing a significant difference in the taskwork workload scores between the 0- and 2-back tasks ($p < 0.001$). Thus, the 2-back task did serve as the more demanding secondary task than the 0-back task.

For testing **H1**, we performed k-shape clustering using the profiles of human key input ratio (i.e., DV1) across the four levels (i.e., the agent capabilities and the n-back task difficulty levels), and Figure 10 presents the relationships between the generated clusters and game scores across the four levels. We ran one-way ANOVAs to compare the clusters' game scores across the four levels. The ANOVAs revealed significant differences in the game scores between the clusters in

[10]Although our target sample size was 192 as we preregistered, there were participants who completed the familiarization session and Block 1, but did not resume and complete the entire experiment, resulting in the smaller total number.

the more capable agent with the 0-back task ($F_{3,732}$ = 11.3, p < 0.001, η_p^2 = 0.0443) and the 2-back task ($F_{3,728}$ = 16.1, p < 0.001, η_p^2 = 0.0623) and in the less capable with the 2-back task ($F_{3,634}$ = 9.70, p < 0.001, η_p^2 = 0.0439); while no differences were confirmed in the less capable with with the 0-back task ($F_{3,642}$ = 1.99, p = 0.114, η_p^2 = 0.00921). The follow-up analyses showed significant differences between Cluster 1 and other three clusters in the more capable agent with the 0- and 2-back tasks and the less capable with the 2-back task (p < 0.01). The results suggest that high-performing teams exhibited different profiles of human key input ratio when compared to lower-performing teams although such trends were observed only in the more capable agent condition, and therefore, we partially confirmed **H1**.

For investigating **H2a**, we carried out chi-square tests for the more capable agent with the 0- & 2-back tasks and the less capable with the 2-back task condition, detecting significant differences in the UI design frequencies between the clusters; the more capable with the 0-back ($\chi^2(6)$ = 172, p < 0.001), the 2-back ($\chi^2(6)$ = 130, p < 0.001), and the less capable agent with the 2-back task ($\chi^2(6)$ = 77.8, p < 0.001). Follow-up analyses were done using Pearson residuals, and Figure 11 shows mosaic plots, highlighting in which clusters each UI exhibited a greater and/or a smaller frequency than the expected values. We expected the agent state UI design to appear more dominantly in the high-performing cluster; however, surprisingly, the baseline design exhibited such a pattern (in the more capable agent condition). Therefore, the results did not support **H2a**.

To examine **H2b** and **H2c**, we performed a three-way MANOVA, indicating a significant interaction effect between the agent capability and UI design ($MATS$ = 48.7, p < 0.01) while also detecting significant main effects of the agent capability ($MATS$ = 43.6, p < 0.01) and the n-back task level ($MATS$ = 46.2, p < 0.001). As follow-up analyzes, we ran two-way ANOVAs for the overall game scores, teamwork, and team-task balancing workload scores. We confirmed a significant interaction effect between the agent capability and UI design on the overall game scores ($F_{2,338}$ = 16.6, p < 0.001, η_p^2 = 0.0894). Pairwise comparisons showed, in the more capable agent condition, a significant difference between the baseline and lander physics UI designs (p < 0.001), but no differences between the baseline and agent state UI designs (p = 0.0603). As shown in Figure 9a, in the more capable agent condition, the baseline design exhibited higher overall game scores and outperformed the lander physics UI. However, this trend is flipped in the less capable agent condition. The pairwise comparisons showed a significant difference between the baseline and lander physics UI in the less capable agent condition (p < 0.05) whereas no differences between the baseline and agent state were observed (p = 0.180). This means that the lander physics UI outperformed the baseline design in the less capable agent condition, but the agent state UI did not contrary to our expectations. Thus, we did not corroborate **H2b**.

As for the workload measures, we found significant main effects of the agent capability ($F_{1,338}$ = 6.25, p < 0.05, η_p^2 = 0.0181) and UI design ($F_{2,338}$ = 3.23, p < 0.05, η_p^2 = 0.0187) on the team-task balancing workload scores whereas only a significant main effect of the agent capability on the teamwork workload was observed ($F_{2,338}$ = 7.99, p < 0.01, η_p^2 = 0.0231). To investigate where the difference in the team-task balancing workload between the three UIs lies, we performed pairwise comparisons, suggesting a significant difference between the baseline and lander physics UI design (p < 0.05); there were no significant differences between the agent state UI and either the lander physics UI (p = 0.203) or the baseline design (p = 1.00). As shown in Figure 9d, the participants in the group of the lander physics were likely to report higher team-task workload scores when compared to the baseline group. With these results, we did not confirm **H2c** either.

6.2 Subjective Feedback

To facilitate our discussion, we list some subjective feedback provided by the participants during the debriefing session in Table 2.

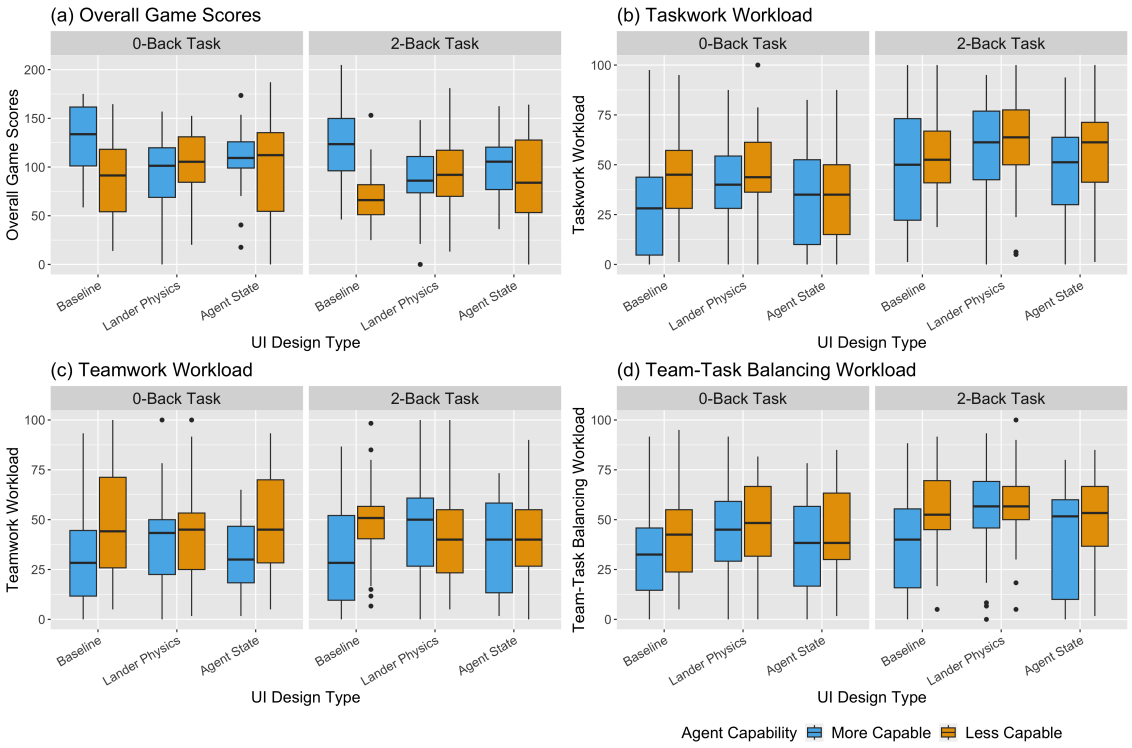


Fig. 9. (a) The baseline design outperformed the lander physics UI design in the more capable agent condition ($p < 0.001$) whereas we did not detect any significant differences between the baseline and agent state UI designs ($p = 0.0603$). However, we observed the opposite trend in the case of the less capable agent condition, meaning that the lander physics outperformed the baseline design ($p < 0.05$); the relationship between the baseline and agent state UI designs remain same ($p = 0.180$). (b) The taskwork workload scores were used for the manipulation check, suggesting a significant difference between the 0- and 2-back tasks ($p < 0.001$). (c) We detected a main effect of the agent capability was found ($p < 0.01$). (d) In the less capable agent condition, the lander physics UI design exhibited higher team-task balancing scores than the baseline design ($p < 0.05$) whereas no significant differences were detected between the agent state and either the baseline ($p = 1.00$) or the lander physics ($p = 0.203$).

7 DISCUSSION

7.1 Understanding Patterns of Interactions for More Informed Decision on UI Design

We expected the agent state UI design to allow humans to interact with the agents with a distinguishable pattern, resulting in higher game scores. However, surprisingly, such a pattern was observed with the baseline design UI design in the more capable agent condition, and the baseline design outperformed the lander physics UI design. Our demographic questionnaire responses did not show any skewed distributions of the participants in terms of their game experience across the levels. We believe that the additional UI features of the lander physics design were prone to induce humans to actively, but unnecessarily participate in the maneuvering task during the first flight phase, leading to less efficient maneuvering. The profiles of Cluster 1 in Figures 10a and b suggest that actively participating in the maneuvering task right before landing enabled higher game scores when working with the more capable agent. This trend was also confirmed by the subjective feedback from P31 and P96 (Table 2).

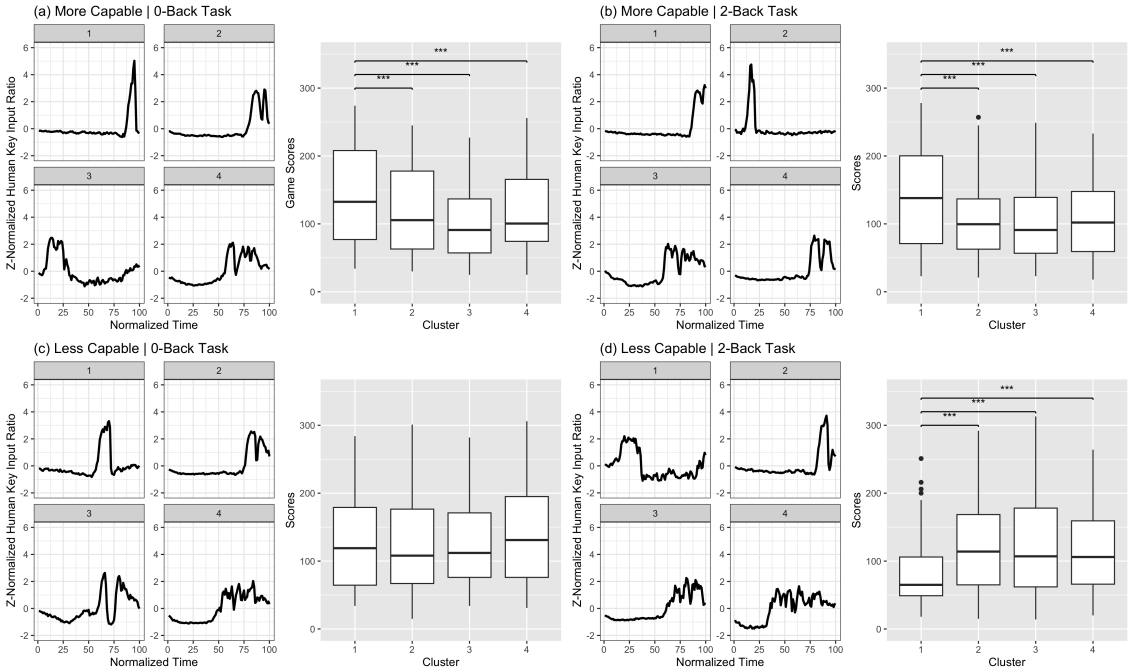


Fig. 10. The line plots show the relationships between the normalized time vs. z-normalized human key input ratio; each line depicts a centroid of profiles in its cluster. The box plots show game scores in each cluster; *** indicates $p < 0.001$. (a) Cluster 1 exhibited higher game scores than Clusters 2-4, and therefore it is considered to be a high-performing cluster. Trials in Cluster 1 tend to show less human interactions in the first phase of flight. (b) Likewise, Cluster 1 is a high-performing cluster given higher game scores than the other three clusters. Cluster 1 contains trials that show less human interactions in the first phase of flight. (c) There were no significant differences between the four clusters. (d) Cluster 1 showed lower game scores than Clusters 2-4, suggesting that Cluster 1 should be a low-performing cluster. The line plot of Cluster 1 indicates that more human interactions in the first phase of flight are prone to undermine game scores.

However, in the less capable agent condition, the relationship between the baseline and lander physics UI designs was flipped, meaning that the baseline design underperformed the lander physics UI design. The baseline design was more susceptible to differences in agent capability, resulting in a significant decrease in the overall game scores in the less capable agent condition (see Figure 9a). This observation stresses the importance of providing information about a compositional control collaborative task to humans working with a low capable agent. As seen in Figure 10d, Cluster 1 (i.e., the low-performing cluster in which the baseline design was more likely to appear than the expected frequency) showed a certain amount of human interactions in the first phase, but less participation at the end of the landing attempt; the similar profile was found in Cluster 3 in Figure 10a. In contrast, the opposite trend was observed in the other three clusters.

In the experiment, the agent state UI did not exhibit specific profiles of human key input ratio leading to game scores superior to the other two UI designs, which did not corroborate our hypotheses. Still, our approach to focusing on interaction patterns helps us understand characteristics exhibited by high- and low-performing teams, offering additional design insights. For instance, in the case of this game experiment setting, our observation in the relationship between the baseline and lander physics UI designs informed us of the need for avoiding human over-participation in

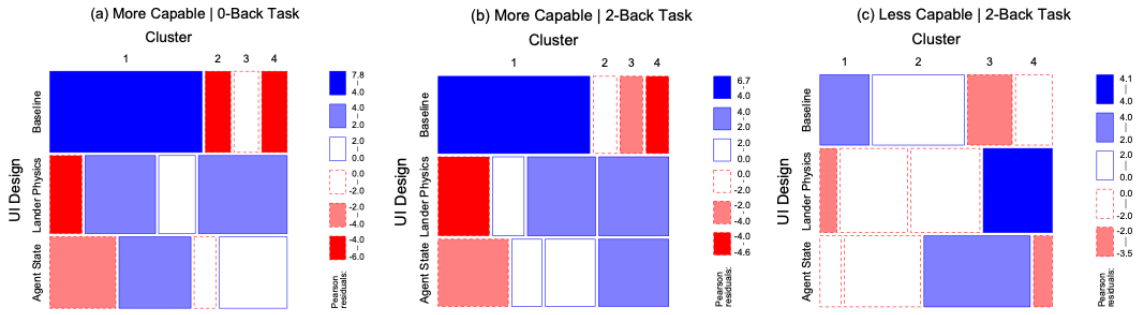


Fig. 11. These mosaic plots represent how more or less likely each UI was to be observed in each cluster than the expected values. Cold colors suggest a trend where a UI was more likely to appear than the expected value. In contrast, cells are filled out with warm colors when a UI was less likely to appear than the expected value. (a) & (b) The baseline UI design exhibited a greater frequencies in Cluster 1 (i.e., the high-performing cluster) whereas the other two UI designs were less likely observed in Cluster 1. (c) The baseline UI design tended to appear in Cluster 1 (i.e., the low-performing cluster) in the less capable agent and the 2-back task condition whereas the lander physics exhibited smaller frequencies in Cluster 1.

Table 2. Subjective feedback on UI designs provided by the participants in the debriefing session. The participants were asked to share their experience with the assigned UI design and type their responses in a text entry box.

ID	Agent	UI	Feedback Type	Quote
P1	M	A	Experience	*"It was hard to understand why there was a successful landing or not a successful landing"*
P29	L	A	Experience	*"The UI was intuitive and easy to understand."*
P31	M	B	Experience	*"The AI made it quite easy but also they did all of the work, you barely had to click any button to move the Moonlander the AI did everything on its own."*
P32	L	L	Recommendation	*"Making the A.I. feature more easy to understand/ anticipate."*
P84	M	L	Recommendation	*"Better visual for AI participation"*
P93	L	L	Recommendation	*"An explanation of what the AI will do during the landing process would be helpful so the user knows what to focus on and when."*
P96	M	B	Experience	*"I didn't have to do the maneuver part at all."*
P100	L	L	Recommendation	*"Highlight when the AI is controlling the spacecraft on the trajectory so it's easier to know when it is primarily guiding and you just need to monitor."*
P104	M	B	Recommendation	*"Allow the user to turn off the engines for the AI."*
P135	L	B	Experience	*"I felt like the AI would fight my input if it was doing something incorrect and I would get distracted from the checkpoints [checklists] trying to correct it."*
P145	M	A	Experience	*"The circle about nervous or confidence helped to control more easier to control speed."*

Agent - M: More, L: Less | **UI** - B: Baseline, L: Lander Physics, A: Agent State

the first flight phase and facilitating their participation before landing. The importance of facilitating participation before landing is more pronounced when humans work with the less capable agent.

7.2 Design Implications for Teamwork-Oriented UI Design

The agent state UI was designed to improve human-agent teamwork presenting information about when the agent needed help from humans in a real-time manner. Although we did not confirm **H2a-c** centered on the agent state UI design, the results offered some implications for designing teamwork-oriented UI. First, overall, the participants tended to accept the agent state UI design features as P29 and P145 articulated (Table 2). Furthermore, the subjective feedback on the lander physics UI appears to reinforce the consideration of a teamwork-oriented UI design. As P32, P84, P93, and P100 hinted, humans seek information about what the agent is currently doing (P84 & P100) and will do in the near future (P32 & P93), conforming the SAT model [13, 14] and observability and predictability in the interdependence analysis [48].

In addition to the subjective feedback, the results of the team-task balancing workload scores (Figure 9d) bolster the potential of the agent state UI design to improve human-agent team performance. The participants using the lander physics UI design were prone to indicate higher team-task balancing workload scores than the baseline UI design. In contrast, we did not observe significant differences in the team-task balancing workload scores between the agent state and baseline UI designs. Indeed, the agent state UI design did not exhibit significantly higher game scores than the baseline in the either more or less capable agent condition. However, the results suggested that adding the agent state UI features did not lead to a significant increase in the team-task balancing workload scores, which appears to be promising. The next obvious step is to refine the agent state UI features to improve overall team performance without taxing humans' workload, which remains an open question.

With respect to the agent state UI design, the nature of how the more and less capable agents were developed may have blurred the expected performance level. We employed the notion of agent's malfunction by introducing the probabilities of taking no actions over the course of the landing attempt. This approach, indeed, induced deviations from the agent's intended flight path and vertical speed. However, the agents were able to relatively quickly recover from the deviated state independently (i.e., the agent expressed its high confidence level shortly after requesting human's need). Although we did not observe a significant increase in either the teamwork or team-task balancing workload scores, this agent behavior may have confused the participants and diminished the expected performance of the agent state UI design, which is underpinned by the subjective feedback from P1 (Table 2). We believe that there would be a possibility to obtain different results employing different types of agents, including, for instance, an agent failing to maneuver in a specific context rather than introducing malfunctions over the course of a landing attempt. This will play an important role in upcoming experiments.

In the experiment, we did not discover conclusive evidence showing that the agent state UI design improved overall game scores. Yet, it is worthwhile further exploring the design space of teamwork-oriented UI allowing better team performance considering the subjective feedback and the team-task balancing workload scores. This is supported by the self-driving simulation study results reported by Peintner et al. [76] finding that not only is it important for UI designs to convey confidence levels, but it is important that they provide humans with an option to work together with agents and to influence agents' behavior.

7.3 Input Composition Function in Compositional Control Collaboration

The moon lander maneuvering task was presented as a compositional control collaborative task, where both the player and the agent blended their inputs, and the combined inputs affected the system (i.e., the moon lander). The experiment done by Momose et al. [70] reported a trend where the participants did not change a control authority level between a

human and an agent and tended to work in a fifty-fifty control authority configuration (i.e., a default setting) in their compositional control collaboration task setting. Based on their observation, we did not offer any capabilities to change the control authority level in the experiment. However, the subjective feedback from a few participants (P104 & P135) implied the need for offering a way to determine how to compose inputs from the player and agent in a flexible, timely manner. Even with the more capable agent that was designed to exhibit high landing performance levels, humans still appear to want a certain degree of control, which is in alignment with findings by Roy et al. [83].

One of the easiest ways to do this would be, as P104 suggested, to enable a manual control configuration by providing an agent disengage option to humans. Another approach would be to implement a composition function where more control authority levels are given to humans as more conflicts occur; in this case, an agent would be designed to implicitly understand human's intent and explicitly grant more control authority to humans. Due to the fact that inputs are continuous, there are myriad of ways to compose inputs from team members. Therefore, this opens another research area to investigate the best approach to composing inputs to improve team performance in a compositional control collaboration team setting.

7.4 Limitations and Future Work

We acknowledged several limitations in the experiment. First, we could not hit the target sample size; we excluded the four participants as well as had 15 participants who have not completed the entire experiment. Also, the study participants were primarily young coming from the university community; it would be useful to acquire a more diverse population in a future investigation. Yet, our demographic questionnaire shows a wide range of participant's game experience, and therefore we believe that game experience did not significantly confound the results.

In the experiment, we administered the modified version of the TWLQ, offering the preliminary insights into the team-task balancing workload across the three UI designs. Yet, the questionnaire validity needs to be checked, and each question item should be refined so that we could better examine differences in the teamwork and the team-task balancing workload between UI designs.

For testing **H1** and **H2a**, we employed k-shape clustering and pre-specified four clusters because we expected the four quartiles to be sufficient and wanted to avoid proliferating the number of clusters. However, just four clusters may potentially make it more difficult to draw a conclusion. In future experiments, we will consider different approaches so that we could gain a clearer insight into the relationship between patterns of interactions and UI designs in future work.

As discussed above, the agent implementation in the experiment may have diluted the expected performance of the agent state UI design. The UI showed the agent request for help, but based on the agent's ability to recover back to the intended flight path, the UI changed back to normal, which may have confused some participants using the agent state UI design.

One condition that was not tested in this experiment was a complete takeover by the participant for the agent, which happens when the agent reaches a failure mode. A non-functional AI teammate would force the participant to rapidly assess situation awareness and then respond to complete a successful landing. Our hypothesis is that the agent state UI would allow participants to achieve higher performance by letting the participant know when the agent was working well and when it needed help. The goal of showing trends in nervousness and the urgency of the need for help is to enable the participant to anticipate the need for help and to provide that help early, before a takeover request event can occur. Takeover requests such as these are common in self-driving cars and with autopilots for air and sea vehicles. Expanding the testing envelope to include full takeover may highlight an untested capability of the agent state UI: allowing the operator to anticipate and course correct before the takeover request event occurs.

The experiment employed the compositional control collaboration team setting, where only two team members worked together (i.e., dyad). Although near-term real-world applications of compositional control collaboration teams are anticipated to be such a dyad type of teams (e.g., a human driver and an automated driving agent in self-driving, a doctor and AI in medical decision-making), it is interesting to scrutinize compositional control collaboration teams involving three or more entities. Such further investigations are expected to identify additional interesting dimensions to distinguish different types of teams, helping to polish our proposed approach to classifying teams.

We believe that our approach to mapping out different types of teams serves as a springboard for advancing research on human-AI teaming by facilitating a more consistent use of the 3Cs. Also, our lessons learned from the experiment are expected to be transferable to more realistic compositional control collaboration team settings, including an automated driving scenario. One of the legal and ethical issues with self-driving lies in "hyping" its driving capability [26, 55], inducing people to possess inappropriate reliance. The driving simulator study by Peintner et al. [76] implied the need for offering a way to human drivers to collaborate with a self-driving agent even in SAE Level 4 or 5. We expect teamwork-oriented UIs to help human drivers achieve appropriate reliance on a self-driving agent, enabling safer driving.

8 CONCLUSION

We offered three contributions in this paper. First, we proposed the new approach to distinguishing types of teams focusing on input compositionality. We introduced the notion of compositional control teams, where all team members possess the same action channels simultaneously, and the system is affected by composed inputs from all team members. Automated driving and human-AI decision-making tasks fall into this type of team. Second, we discussed the relationship between input compositionality and the 3Cs aiming at a more consistent use of the 3Cs. Our approach to mapping out different types of teamwork sheds the light on compositional control collaborative teams as such teams are a unique, interesting type of team. Therefore, we conducted the experiment using the moon lander game to gain insights into UI designs to amplify human-AI team performance in the compositional control collaborative team setting. The results stressed the importance of offering information about the compositional control collaborative task, in particular, in a case where humans work with a less capable agent. Also, while further investigations are certainly required, the results hinted that conveying in a real-time manner the degree to which an agent needs help from humans could improve team performance without significantly taxing in humans' workload. In future work, we continue to examine teamwork-oriented UI design in the context of compositional control collaboration human-agent teaming and aim to apply design insights to more real-world compositional control collaboration team settings, including AI-enabled self-driving cars.

REFERENCES

[1] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. *CoRR* abs/2001.09977 (2020). arXiv:2001.09977 https://arxiv.org/abs/2001.09977

[2] National Highway Traffic Safety Administration. 2022. Levels of Automation. https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-05/Level-of-Automation-052522-tag.pdf Last accessed 9 October 2023.

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[4] Wendy L Bedwell, Jessica L Wildman, Deborah DiazGranados, Maritza Salazar, William S Kramer, and Eduardo Salas. 2012. Collaboration at work: An integrative multilevel conceptualization. *Human resource management review* 22, 2 (2012), 128–145.

[5] Johannes Beller, Matthias Heesen, and Mark Vollrath. 2013. Improving the driver–automation interaction: An approach using automation uncertainty. *Human factors* 55, 6 (2013), 1130–1141.
[6] Ralph Bergmüller, Rufus A Johnstone, Andrew F Russell, and Redouan Bshary. 2007. Integrating cooperative breeding into theoretical concepts of cooperation. *Behavioural processes* 76, 2 (2007), 61–72.
[7] Luzheng Bi, Cuntai Guan, et al. 2019. A review on EMG-based motor intention prediction of continuous human upper limb motion for human-robot collaboration. *Biomedical Signal Processing and Control* 51 (2019), 113–127.
[8] Kristin E Bonnie and Frans BM de Waal. 2004. 11 Primate Social Reciprocity and the Origin of Gratitude. *The psychology of gratitude* (2004), 213.
[9] Clint A Bowers, Curt C Braun, and Ben B Morgan Jr. 1997. Team workload: Its meaning and measurement. In *Team performance assessment and measurement*. Psychology Press, 97–120.
[10] Abhizna Butchibabu, Christopher Sparano-Huiban, Liz Sonenberg, and Julie Shah. 2016. Implicit coordination strategies for effective team communication. *Human factors* 58, 4 (2016), 595–610.
[11] Tara Capel and Margot Brereton. 2023. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–23.
[12] Xavier Castañer and Nuno Oliveira. 2020. Collaboration, coordination, and cooperation among organizations: Establishing the distinctive meanings of these terms through a systematic literature review. *Journal of Management* 46, 6 (2020), 965–1001.
[13] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science* 19, 3 (2018), 259–282.
[14] Jessie Y Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael Barnes. 2014. Situation awareness-based agent transparency. *US Army Research Laboratory* (2014), 1–29.
[15] Erin K Chiou, Mustafa Demir, Verica Buchanan, Christopher C Corral, Mica R Endsley, Glenn J Lematta, Nancy J Cooke, and Nathan J McNeese. 2021. Towards Human–Robot Teaming: Tradeoffs of Explanation-Based Communication Strategies in a Virtual Search and Rescue Task. *International Journal of Social Robotics* (2021), 1–20.
[16] Erin K Chiou, John D Lee, and Tianshuo Su. 2019. Negotiated and reciprocal exchange structures in human-agent cooperation. *Computers in Human Behavior* 90 (2019), 288–297.
[17] Andrzej Cichocki and Alexander P Kuleshov. 2021. Future trends for human-ai collaboration: A comprehensive taxonomy of AI/AGI Using Multiple Intelligences and Learning Styles. *Computational Intelligence and Neuroscience* 2021 (2021), 1–21.
[18] Nazli Cila, Iskander Smit, Elisa Giaccardi, and Ben Kröse. 2017. Products as agents: Metaphors for designing the products of the IoT age. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 448–459.
[19] Mark S Copelovitch and Tonya L Putnam. 2014. Design in context: Existing international agreements and new cooperation. *International Organization* 68, 2 (2014), 471–493.
[20] Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. 2023. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International* (2023), 1–12.
[21] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. 2020. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630* (2020).
[22] Thomas H Davenport and Nitin Mittal. 2022. How generative AI is changing creative work. *Harvard Business Review* (2022).
[23] David De Cremer and Garry Kasparov. 2021. AI should augment human intelligence, not replace it. *Harvard Business Review* 18 (2021), 1.
[24] G-J De Vreede and Robert O Briggs. 2005. Collaboration engineering: designing repeatable processes for high-value collaborative tasks. In *Proceedings of the 38th annual Hawaii international conference on system sciences*. IEEE, 17c–17c.
[25] Mustafa Demir, Aaron D Likens, Nancy J Cooke, Polemnia G Amazeen, and Nathan J McNeese. 2018. Team coordination and effectiveness in human-autonomy teaming. *IEEE Transactions on Human-Machine Systems* 49, 2 (2018), 150–159.
[26] Liza Dixon. 2020. Autonowashing: The greenwashing of vehicle automation. *Transportation research interdisciplinary perspectives* 5 (2020), 100113.
[27] Damian Okaibedi Eke. 2023. ChatGPT and the rise of generative AI: threat to academic integrity? *Journal of Responsible Technology* 13 (2023), 100060.
[28] Lisa A Elkin, Matthew Kay, James J Higgins, and Jacob O Wobbrock. 2021. An aligned rank transform procedure for multifactor contrast tests. In *The 34th annual ACM symposium on user interface software and technology*. 754–768.
[29] Mica R Endsley. 2023. Supporting Human-AI Teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior* 140 (2023), 107574.
[30] Mica R Endsley, Betty Bolté, and Debra G Jones. 2003. *Designing for situation awareness: An approach to user-centered design*. CRC press.
[31] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. 2023. Art and the science of generative AI. *Science* 380, 6650 (2023), 1110–1111.
[32] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* 378, 6624 (2022), 1067–1074.
[33] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[34] Carlos Ferran-Urdaneta. 1999. Teams or communities? Organizational structures for knowledge management. In *Proceedings of the 1999 ACM SIGCPR conference on computer Personnel Research*. 128–134.

[35] Frank Flemisch, Johann Kelsch, Christan Löper, Anna Schieben, Julian Schindler, and Matthias Heesen. 2008. Cooperative control and active interfaces for vehicle assitsance and automation. (2008).

[36] Sarah Friedrich, Frank Konietschke, and Markus Pauly. 2019. Resampling-based analysis of multivariate data and repeated measures designs with the R package MANOVA. RM. (2019).

[37] Anne Haugen Gausdal, Helge Svare, and Guido Möllering. 2016. Why don't all high-trust networks achieve strong network benefits? A case-based exploration of cooperation in Norwegian SME networks. *Journal of Trust Research* 6, 2 (2016), 194–212.

[38] Riccardo Gervasi, Luca Mastrogiacomo, and Fiorenzo Franceschini. 2020. A conceptual framework to evaluate human-robot collaboration. *The International Journal of Advanced Manufacturing Technology* 108 (2020), 841–865.

[39] Elaheh Ghasemi, Nadia Lehoux, and Mikael Rönnqvist. 2022. Coordination, cooperation, and collaboration in production-inventory systems: a systematic literature review. *International Journal of Production Research* (2022), 1–32.

[40] Manu Goyal, Thomas Knackstedt, Shaofeng Yan, and Saeed Hassanpour. 2020. Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Computers in biology and medicine* 127 (2020), 104065.

[41] Eric T Greenlee, Gregory J Funke, and Lindsay Rice. 2019. Evaluation of the Team Workload Questionnaire (TWLQ) in a Team-Choice Task. *Human Factors* 61, 2 (2019), 348–359.

[42] Maaike Harbers, Catholijn Jonker, and Birna Van Riemsdijk. 2012. Enhancing team performance through effective communication. In *Proceedings of the 4th Annual Human-Agent-Robot Teamwork Workshop*. 1–2.

[43] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. 2013. Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications*. 210–217.

[44] Eric Holder and Ning Wang. 2021. Explainable artificial intelligence (XAI) interactively working with humans as a junior cyber analyst. *Human-Intelligent Systems Integration* 3, 2 (2021), 139–153.

[45] John R Hollenbeck, Bianca Beersma, and Maartje E Schouten. 2012. Beyond team types and taxonomies: A dimensional scaling conceptualization for team description. *Academy of Management Review* 37, 1 (2012), 82–106.

[46] Mohammad Hossein Jarrahi. 2018. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business horizons* 61, 4 (2018), 577–586.

[47] Matthew Johnson and Jeffrey M Bradshaw. 2021. How Interdependence Explains the World of Teamwork. In *Engineering Artificially Intelligent Systems*. Springer, 122–146.

[48] 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.

[49] Matthew Johnson, Catholijn Jonker, Birna Van Riemsdijk, Paul J Feltovich, and Jeffrey M Bradshaw. 2009. Joint activity testbed: Blocks world for teams (BW4T). In *Engineering Societies in the Agents World X: 10th International Workshop, ESAW 2009, Utrecht, The Netherlands, November 18-20, 2009. Proceedings 10*. Springer, 254–256.

[50] Matthew Johnson and Alonso Vera. 2019. No AI is an island: the case for teaming intelligence. *AI magazine* 40, 1 (2019), 16–28.

[51] Jon R Katzenbach and Douglas K Smith. 2015. *The wisdom of teams: Creating the high-performance organization*. Harvard Business Review Press.

[52] Ari Kolbeinsson, Erik Lagerstedt, and Jessica Lindblom. 2019. Foundation for a classification of collaboration levels for human-robot cooperation in manufacturing. *Production & Manufacturing Research* 7, 1 (2019), 448–471.

[53] Alexander Kott, Paul Théron, Martin Drašar, Edlira Dushku, Benoît LeBlanc, Paul Losiewicz, Alessandro Guarino, Luigi Mancini, Agostino Panico, Mauno Pihelgas, et al. 2018. Autonomous intelligent cyber-defense agent (AICA) reference architecture. Release 2.0. *arXiv preprint arXiv:1803.10664* (2018).

[54] Herbert Kotzab, Inga-Lena Darkow, Ilja Bäumler, and Christoph Georgi. 2019. Coordination, cooperation and collaboration in logistics and supply chains: a bibliometric analysis. *Production* 29 (2019).

[55] Tom Krisher. 2023. As a criminal case against a Tesla driver wraps up, legal and ethical questions on Autopilot endure. https://abcnews.go.com/US/wireStory/criminal-case-tesla-driver-ends-legal-ethical-questions-102274824 Last accessed 9 October 2023.

[56] Alexander Kunze, Stephen J Summerskill, Russell Marshall, and Ashleigh J Filtness. 2019. Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics* 62, 3 (2019), 345–360.

[57] Shan G Lakhmani, Julia L Wright, Michael R Schwartz, and Daniel Barber. 2019. Exploring the effect of communication patterns and transparency on performance in a human-robot team. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 160–164.

[58] Carl Larson, Carl E Larson, and Frank MJ LaFasto. 1989. *Teamwork: What must go right/what can go wrong*. Vol. 10. Sage.

[59] Joonbum Lee, Hansol Rheem, John D Lee, Joseph F Szczerba, and Omer Tsimhoni. 2023. Teaming with your car: Redefining the driver–automation relationship in highly automated vehicles. *Journal of cognitive engineering and decision making* 17, 1 (2023), 49–74.

[60] Leonidas C Leonidou, Saeed Samiee, Bilge Aykol, and Michael A Talias. 2014. Antecedents and outcomes of exporter–importer relationship quality: synthesis, meta-analysis, and directions for further research. *Journal of international marketing* 22, 2 (2014), 21–46.

[61] Laura Lucaj, Patrick van der Smagt, and Djalel Benbouzid. 2023. AI Regulation Is (not) All You Need. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1267–1279.

[62] Joseph B Lyons, Katia Sycara, Michael Lewis, and August Capiola. 2021. Human–autonomy teaming: Definitions, debates, and directions. *Frontiers in Psychology* 12 (2021), 589585.

[63] Thomas W Malone and Kevin Crowston. 1990. What is coordination theory and how can it help design cooperative work systems?. In *Proceedings of the 1990 ACM conference on Computer-supported cooperative work*. 357–370.

[64] 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.

[65] Eric Martin, Isabelle Nolte, and Emma Vitolo. 2016. The Four Cs of disaster partnering: Communication, cooperation, coordination and collaboration. *Disasters* 40, 4 (2016), 621–643.

[66] Laura McGirr, Yan Jin, Mark Price, Andrew West, Katherine van Lopik, and Vincent McKenna. 2022. Human Robot Collaboration: Taxonomy of Interaction Levels in Manufacturing. In *ISR Europe 2022; 54th International Symposium on Robotics*. VDE, 1–8.

[67] Madeleine McNamara. 2012. Starting to untangle the web of cooperation, coordination, and collaboration: A framework for public managers. *International Journal of Public Administration* 35, 6 (2012), 389–401.

[68] Nathan J McNeese, Mustafa Demir, Nancy J Cooke, and Christopher Myers. 2018. Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human factors* 60, 2 (2018), 262–273.

[69] Kazuhiko Momose, Rahul Mehta, Josias Moukpe, Troy Weekes, Thomas Eskridge, and Daniel Kidwell. 2023. Compositional Human-Agent Teams for Spaceflight. (2023).

[70] Kazuhiko Momose, Troy Weekes, Rahul Mehta, Cameron Wright, Josias Moukpe, and Thomas Eskridge. 2023. Patterns of Effective Human-Agent Teams. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[71] Jakob Nielsen. 1994. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 152–158.

[72] Helle Aarøe Nissen, Majbritt Rostgaard Evald, and Ann Højbjerg Clarke. 2014. Knowledge sharing in heterogeneous teams through collaboration and cooperation: Exemplified through Public–Private-Innovation partnerships. *Industrial Marketing Management* 43, 3 (2014), 473–482.

[73] Donald A Norman. 1988. *The psychology of everyday things*. Basic books.

[74] Society of Automotive Engineers (SAE) International. 2021. SAE Levels of Driving Automation™ Refined for Clarity and International Audience. https://www.sae.org/binaries/content/assets/cm/content/blog/sae-j3016-visual-chart_5.3.21.pdf Last accessed 9 October 2023.

[75] Adrian M Owen, Kathryn M McMillan, Angela R Laird, and Ed Bullmore. 2005. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping* 25, 1 (2005), 46–59.

[76] Jakob Peintner, Carina Manger, and Andreas Riener. 2023. Communication of Uncertainty Information in Cooperative, Automated Driving: A Comparative Study of Different Modalities. In *15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 322–332.

[77] Jakob Benedikt Peintner, Carina Manger, and Andreas Riener. 2022. "Can you rely on me?" Evaluating a Confidence HMI for Cooperative, Automated Driving. In *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 340–348.

[78] Alex Sandy Pentland. 2012. The new science of building great teams. *Harvard business review* 90, 4 (2012), 60–69.

[79] Elizabeth Phillips, Kristin E Schaefer, Deborah R Billings, Florian Jentsch, and Peter A Hancock. 2016. Human-animal teams as an analog for future human-robot teams: Influencing design and fostering trust. *Journal of Human-Robot Interaction* 5, 1 (2016), 100–125.

[80] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 379–396.

[81] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. 2022. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific reports* 12, 1 (2022), 14952.

[82] Jennifer Rix. 2022. From Tools to Teammates: Conceptualizing Humans' Perception of Machines as Teammates with a Systematic Literature Review. (2022).

[83] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. Automation accuracy is good, but high controllability may be better. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.

[84] Maria Jesus Saenz, Elena Revilla, and Cristina Simón. 2020. Designing AI systems with human-machine teams. *MIT Sloan Management Review* 61, 3 (2020), 1–5.

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

[86] Eduardo Salas, Nancy J Cooke, and Michael A Rosen. 2008. On teams, teamwork, and team performance: Discoveries and developments. *Human factors* 50, 3 (2008), 540–547.

[87] Eduardo Salas, Terry L Dickinson, Sharolyn A Converse, and Scott I Tannenbaum. 1992. Toward an understanding of team performance and training. (1992).

[88] Alexis Sardá-Espinosa. 2017. Comparing time-series clustering algorithms in r using the dtwclust package. *R package vignette* 12 (2017), 41.

[89] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[90] James Sellers, William S Helton, Katharina Näswall, Gregory J Funke, and Benjamin A Knott. 2014. Development of the team workload questionnaire (TWLQ). In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 58. Sage Publications Sage CA: Los Angeles, CA, 989–993.

[91] James Sellers, William S Helton, Katarina Näswall, Gregory J Funke, and Benjamin A Knott. 2015. The Team Workload Questionnaire (TWLQ) A Simulated Unmanned Vehicle Task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 59. Sage Publications Sage CA: Los Angeles, CA, 1382–1386.

[92] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[93] Ben Shneiderman and Catherine Plaisant. 2010. *Designing the user interface: Strategies for effective human-computer interaction*. Pearson Education India.

[94] Matthew Sidji, Wally Smith, and Melissa J Rogerson. 2023. The Hidden Rules of Hanabi: How Humans Outperform AI Agents. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[95] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.

[96] Ivan Dale Steiner. 1972. *Group process and productivity*. Academic press New York.

[97] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with humans without human data. *Advances in Neural Information Processing Systems* 34 (2021), 14502–14515.

[98] R Core Team et al. 2013. R: A language and environment for statistical computing. (2013).

[99] Jurriaan Van Diggelen, Jeffrey M Bradshaw, Tim Grant, Matthew Johnson, and Mark Neerincx. 2009. Policy-based design of human-machine collaboration in manned space missions. In *2009 Third IEEE International Conference on Space Mission Challenges for Information Technology*. IEEE, 376–383.

[100] Christian Wankmüller and Gerald Reiner. 2020. Coordination, cooperation and collaboration in relief supply chain management. *Journal of Business Economics* 90 (2020), 239–276.

[101] Astrid Weiss, Ann-Kathrin Wortmeier, and Bettina Kubicek. 2021. Cobots in industry 4.0: A roadmap for future practice studies on human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 51, 4 (2021), 335–345.

[102] Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology innovation management review* 9, 11 (2019).

[103] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[104] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330, 6004 (2010), 686–688.

[105] Canjun Yang, Yuanchao Zhu, and Yanhu Chen. 2021. A review of human–machine cooperation in the robotics domain. *IEEE Transactions on Human-Machine Systems* 52, 1 (2021), 12–25.

[106] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering* 2, 10 (2018), 719–731.

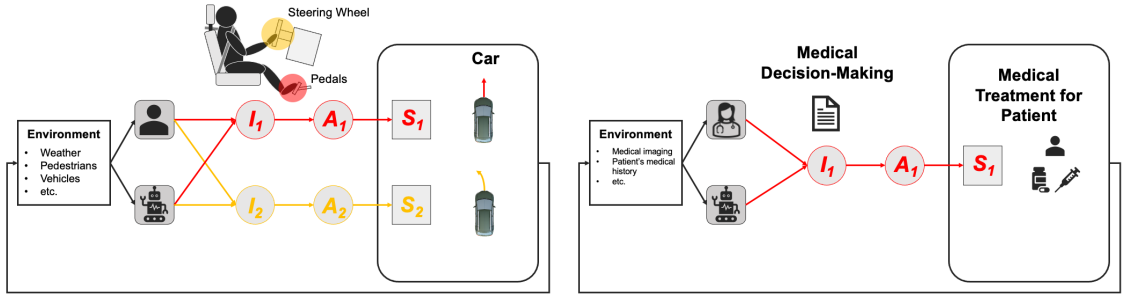
[107] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

[108] Ninger Zhou, Lorraine Kisselburgh, Senthil Chandrasegaran, S Karthik Badam, Niklas Elmqvist, and Karthik Ramani. 2020. Using social interaction trace data and context to predict collaboration quality and creative fluency in collaborative design learning environments. *International Journal of Human-Computer Studies* 136 (2020), 102378.

[109] Yu-Qian Zhu, Donald G Gardner, and Houn-Gee Chen. 2018. Relationships between work team climate, individual motivation, and creativity. *Journal of Management* 44, 5 (2018), 2094–2115.

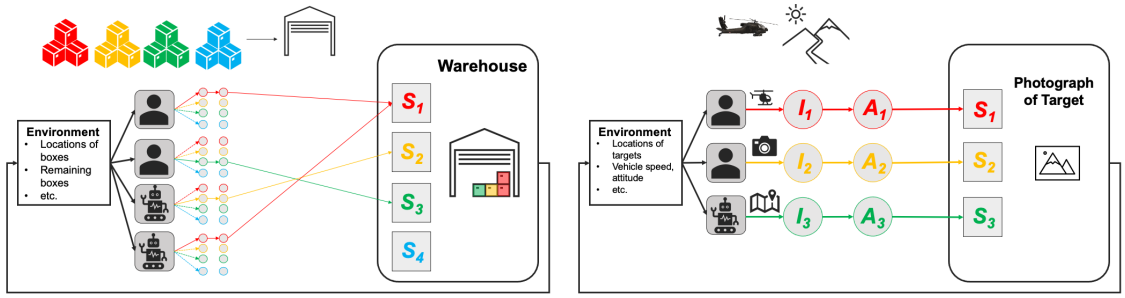
A EXAMPLES OF COMPOSITONAL AND NON-COMPOSITIONAL CONTROL TEAMS

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009



(a) Composition Control Collaboration (Continuous) (b) Composition Control Collaboration (Discrete)

Fig. 12. Examples of compositional control teams. (a) In a self-driving car [2, 74], a human driver and an agent have the same action channels; a steering wheel and gas pedals. One of the examples of compositional control collaboration is "Horse-Mode" [35], where an automated system controls a vehicle while a human driver can apply his or her inputs. (b) In a collaborative medical decision-making (e.g., an AI-assisted image diagnosis [40]), a doctor and an agent generate their own decisions first and then compose them to produce one single final answer.



(a) Non-Composition Control Collaboration (b) Non-Composition Control Coordination

Fig. 13. Examples of non-compositional control teams. (a) In Blocks World for Teams (BW4T) [49], team members are asked to pick boxes up, carry them to a designated location, and drop them off. All team members have the same action capabilities, and each contribution affects the system independently. (b) In a simulation environment used in a study by McNeese et al. [68], three unique roles are assigned to team members, namely a pilot, a photographer, and a navigator. The team's ultimate goal is to efficiently take photographs of targets, requiring coordination.