Teaming Up with Artificial Agents in Non-routine Analytical Tasks

LORENZO COMINELLI and FEDERICO GALATOLO, Department of Information Engineering, University of Pisa, Pisa, Italy
CATERINA GIANNETTI, Department of Economics and Management, University of Pisa, Pisa, Italy and Centro E. Piaggio, University of Pisa, Pisa, Italy
FELICE DELL'ORLETTA and CRISTIANO CIACCIO, Antionio Zampolli Institute of Computational Linguistics National Research Council, Pisa, Italy
PHILIPP CHAPKOVSKI, Institut fur Politikwissenschaft, Universitat Duisburg Essen, Duisburg, Germany
GIULIA VENTURI, Antionio Zampolli Institute of computational linguistics National Research Council, Pisa, Italy

Although AI systems are becoming increasingly common in the workplace, research on their integration into human teams remains limited. In particular, little is known about how the embodiment of artificial agents shapes collaboration and performance in non-routine analytical tasks. To address this gap, we examine how different degrees of embodiment affect team performance and conversational dynamics in a real-life escape room. Teams composed of either three humans or two humans and an artificial agent (a Box, an Avatar, or a hyper-realistic humanoid) worked together to escape the room within a time limit. Our findings show that artificial agents have an uneven impact on team outcomes, with some mixed human–AI teams performing exceptionally well and others markedly worse. Human-only teams, by contrast, display more consistent performance: they are more likely to complete all tasks successfully, although they take longer and commit more errors. We also document a suggestive non-linear relationship between embodiment and team performance. Teams interacting with more embodied agents display conversational patterns that more closely resemble human–human dialogue. Together, these findings show that embodied AI shapes collaboration in complex ways, reinforcing evidence that social cues critically guide teamwork dynamics.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Computing methodologies** → *Intelligent agents*; *Natural language processing*;

Additional Key Words and Phrases: human-AI collaboration, human-robot interaction, embodied AI, team performance, conversational dynamics, non-routine analytical tasks

This work was supported by the Ministero dell'Università e della Ricerca, under the PRIN project "Teaming-up with social artificial agents" (grant no. 2022ALBSWX).

Authors' Contact Information: Lorenzo Cominelli, Department of Information Engineering, University of Pisa, Pisa, Italy; e-mail: lorenzo.cominelli@unipi.it; Federico Galatolo, Department of Information Engineering, University of Pisa, Pisa, Italy; e-mail: federico.galatolo@unipi.it; Caterina Giannetti (corresponding author), Department of Economics and Management, University of Pisa, Pisa, Italy and Centro E. Piaggio, University of Pisa, Pisa, Italy; e-mail: caterina.giannetti@unipi.it; Felice Dell'Orletta, Antionio Zampolli Institute of Computational Linguistics National Research Council, Pisa, Italy; e-mail: felice.dellorletta@ilc.cnr.it; Cristiano Ciaccio, Antionio Zampolli Institute of Computational Linguistics National Research Council, Pisa, Italy; e-mail: cristiano.ciaccio@ilc.cnr.it; Philipp Chapkovski, Institut fur Politikwissenschaft, Universitat Duisburg Essen, Duisburg, Germany; e-mail: chapovski@gmail.com; Giulia Venturi, Antionio Zampolli Institute of computational linguistics National Research Council, Pisa, Italy; e-mail: giulia.venturi@ilc.cnr.it.

ACM 2573-9522/2026/7-ART84
https://doi.org/10.1145/3816428

ACM Reference format:
Lorenzo Cominelli, Federico Galatolo, Caterina Giannetti, Felice Dell'Orletta, Cristiano Ciaccio, Philipp Chapkovski, and Giulia Venturi. 2026. Teaming Up with Artificial Agents in Non-routine Analytical Tasks. *ACM Trans. Hum.-Robot Interact.* 15, 4, Article 84 (July 2026), 32 pages.
https://doi.org/10.1145/3816428

1 Introduction

Unlike earlier automation technologies that replaced humans in routine tasks, generative AI presents profound opportunities for jobs requiring creativity, adaptability, and precision [1]. In these types of occupation, rather than functioning as mere tools, AI players can serve as co-creators and partners [5, 10, 64]. Understanding human–AI collaboration is thus crucial to designing AI that complements human workers rather than replacing or misguiding them [55, 10].

Empirical evidence demonstrates that generative AI significantly enhances productivity in structured tasks, such as standardized writing, customer service, and routine programming, particularly benefiting less experienced workers [11, 23, 52 56]. However, its impact on complex, open-ended tasks is less uniform: AI tends to improve efficiency without necessarily fostering creativity [3] and primarily benefits top performers rather than leveling the playing field [54].

This uneven impact, observed mainly in one-to-one human–AI interactions, raises important questions about AI's role in teamwork, as its effectiveness in complex, team-based environments remains unclear. Recent field evidence suggests that generative AI can already replicate some of the benefits traditionally associated with human collaboration. [22] show that individuals supported by AI achieved performance comparable to small teams, bringing together diverse skills and even improving emotional engagement. However, these effects were observed in non-embodied contexts, leaving open whether AI can effectively operate in settings that demand physical presence, multimodal coordination, and dynamic social interaction, domains where it still seems to struggle [49]. Unlike structured tasks, non-routine teamwork requires AI to support coordination, conflict resolution, and social interaction—areas where current systems face limitations in recognizing social cues, emotional states, and evolving group dynamics [48, 20, 28, 47]. Physical embodiment has been proposed as a way to enhance AI's social presence and engagement in teams [36, 60, 43], but its impact on complex problem-solving remains largely unexplored.

The current study advances this debate on human–AI collaboration by examining how artificial agents influence team performance in the underexplored domain of complex, non-routine analytical tasks. Unlike previous work, we investigate AI's role in multi-agent teamwork, where coordination, adaptability, and social dynamics are crucial. To this end, we leverage a real-life escape room setting, a controlled environment that requires teams to solve intricate problems under time constraints. Prior research has demonstrated the suitability of this type of set-up for studying creativity and problem-solving skills in both real-life settings [32] and digital formats with AI collaborators [3]. Building on this approach, we systematically examine how different AI embodiments—from a simple computer box to an avatar and a humanoid robot—affect team dynamics and productivity. By forming three-player teams with varying compositions (either all-human teams or mixed teams with two humans and an AI player of different embodiments), we assess not only traditional performance metrics, such as escape time and errors, but also the broader impact of AI on team communication and coordination. Our custom-designed escape room consists of four distinct challenges, each targeting a specific skill: linguistic proficiency, logical and mathematical reasoning, historical and geographical knowledge, and problem-solving. This enables a deeper exploration of how AI embodiment influences teamwork beyond structured tasks.

Moreover, we contribute to the recent literature on the broader influence of AI on human communication. Studies on human-only teams indicate that active communication within a team setting enhances decision-making and fosters shared understanding across various types of tasks [2]. Recent studies, however, raise concerns about AI's impact on linguistic diversity and evolving communication patterns [8]. For example, an analysis of academic YouTube videos reveals noticeable shifts in word usage following ChatGPT's release [70]. To investigate these effects in team settings, we analyze conversations between humans who interact with AI of different embodiments using Profiling-**Universal Dependencies (UD)**, a computational stylometry tool [9], and **Sentence-BERT (SBERT)** [57], which provide insights to study the semantic similarity of human conversations with AI agents of different embodiments.

Finally, we complement our analysis by enriching our evidence on performance metrics and communication patterns with additional variables collected through questionnaires. These capture participants' characteristics, perceptions of AI teammates, and self-reported experiences of coordination and engagement. This comprehensive dataset offers a multi-dimensional perspective on the influence of artificial players of different embodiments, contributing to a deeper understanding of the role of artificial agents in complex, team-based problem-solving, all within the context of Italian, an understudied language in this field.

Given the limited empirical evidence on how embodied AI operates in this type of complex team setting, we adopt an exploratory approach rather than preregistering directional hypotheses. We draw on the Uncanny Valley framework only as a broad conceptual motivation [25, 27]: while greater embodiment may increase social presence and facilitate coordination, hyper-realistic agents may also trigger discomfort or cognitive friction. Such effects could manifest not only in team performance but also in the linguistic adaptation and communication patterns that emerge during collaboration. This theoretical ambiguity supports our decision to examine empirically, rather than predict, whether our hyper-realistic humanoid, Abel, crosses such a threshold in complex team tasks.

2 Literature Review

Research on human–AI teaming is still in its early stages, leaving many unanswered questions about how AI affects team communication, coordination, and overall dynamics. Building on insights from **Human–Robot Interaction (HRI)**, recent reviews show that as research moves from dyadic (one-to-one) to group and team settings, it faces both social and experiential challenges, such as understanding group dynamics and social roles, as well as computational and technical ones, including group detection, engagement tracking, and real-time coordination [60, 58, 51]. Much of what we know comes from adapting insights from the literature on human–human team and **Human–Computer Interaction (HCI)** [48, 69]. For instance, research shows that individual attributes, such as gender and personality, influence team performance [46], while team cognition, i.e., the way members think together, share knowledge, and coordinate, plays a critical role in group outcomes [48]. However, these insights do not fully capture the complexities of hybrid teams as human–AI collaboration introduces novel challenges to interpersonal coordination [24, 51]. Recent organizational research has started to address these challenges empirically. In a large-scale field experiment with two-person professional teams, Dell'Acqua et al. [22] show that groups assisted by generative AI achieved higher-quality outcomes on real-world innovation and problem-solving tasks. These teams also reported greater emotional engagement than those without AI support. However, these findings stem from non-embodied, text-based collaboration, leaving open the question of whether similar dynamics emerge in physically co-present, multimodal teamwork.

In particular, successful teams tend to heavily rely on contextual understanding, intent, and nonverbal cues—elements that AI struggles to interpret. This limitation hinders team cognition

and situational awareness, as seen in search-and-rescue simulations where AI's inability to process nonverbal signals led to miscommunication and poor coordination [24]. In technical terms, these limitations reflect persistent gaps in group HRI, where robots still lack robust computational mechanisms for core perception and coordination tasks, such as detecting group structure and shared interaction space, recognizing conversational turns and addressees, and generating context-appropriate behaviours during multiparty exchanges [51]. Nevertheless, research suggests that embodied AI, particularly humanoid robots, can enhance engagement and reduce social distance, making AI teammates feel more natural and intuitive [38, 60].

This shift carries implications for team cohesion, shared understanding, and decision-making processes. Additionally, linguistic adaptation plays a crucial role: individuals often simplify their language and use clearer, more explicit instructions when addressing AI [42], but they also tend to ask for details more often [47]. While this adaptation can enhance human–AI interaction, it may also introduce asymmetries in human-to-human communication styles, potentially disrupting teamwork efficiency.

Recent research highlights that AI communication significantly impacts human trust, coordination, and decision-making. AI teammates that communicate clearly, actively sharing updates and initiating discussions, are perceived as more reliable and effective, whereas passive and erratic AI tend to be overlooked or treated as mere tools [30, 40, 30, 62, 71]. However, excessive communication can lead to cognitive overload and frustration, highlighting the need for a balanced approach to information-sharing [38, 40]. In this regard, studies show that AI with human-like features—such as a face, eyes, and speech-to-text capabilities—and those that express group-based emotions are perceived as more trustworthy, fostering team cohesion [14, 17, 25, 36, 45, 60, 63]. Additionally, adaptive systems that promote balanced participation further enhance group coordination and overall teamwork [19, 53, 63].

From another perspective, economic studies suggest that the presence of an automated teammate can reduce human performance by diminishing social incentives, particularly in tasks where team members perform highly similar roles, such as sequential production tasks [16]. On the other hand, when humans and AI agents perform different tasks, humans tend to increase their effort [37, 44]. The perceived appropriateness of automation for a given task also plays a role: people react negatively to AI involvement in social or "human" tasks but show less resistance in analytical tasks [13]. Similarly, [38] find that perceived competence helps teams leverage AI's knowledge.

Finally, aligning human expectations with the AI agent's actual capabilities is crucial for effective teamwork [6, 24, 48, 72]. When users understand an agent's limitations, collaboration improves, but unrealistic expectations lead to miscommunication and slower performance. Research in economics suggests that individual differences—such as ability and belief calibration—also play a role in how effectively people work with AI. For instance, [11] finds that individuals with higher abilities and well-calibrated confidence benefit more from AI assistance, while those with misaligned expectations may not experience the same productivity gains. However, to our knowledge, no studies have systematically examined how these expectation dynamics play out at the group level. This highlights the importance of expectation management, reinforcing insights from the Robot Social Influence model, which emphasizes that transparent AI behavior and clear communication are key to building trust and enhancing efficiency [34].

To sum up, existing research on human–AI teams highlights challenges that traditional theories of human teamwork and HCI do not fully address. In particular, AI teammates lack social intuition, affect human communication dynamics, struggle with implicit clues, and require humans to set clear expectations about their actions and trust the AI's methods. While we have supporting evidence of AI's impact on human communication and individual performance, we still do not fully understand how AI can communicate effectively in team collaboration. Our study contributes to the literature

by exploring how AI teammates with different levels of embodiment shape coordination and, importantly, alter human communication in a specific setting: non-routine analytical tasks, where AI is expected to have the most significant impact in the near future. Drawing on the Uncanny Valley framework [25, 27], we treat embodiment as a design dimension with potentially non-linear effects: increasing human-likeness may strengthen social presence and coordination, although hyper-realistic agents may elicit discomfort, attentional disruption, or shifts in communicative style. Because these mechanisms remain theoretically unresolved in team contexts, we avoid directional predictions and instead examine their empirical manifestations. By examining the interaction between AI embodiment, team communication dynamics, and performance, we offer new insights into when AI enhances teamwork—and when it disrupts it.

3 Methodology

3.1 Experimental Design

Our experimental design was inspired by [32], who demonstrated that escape-room-type experiments can effectively be used to study non-routine analytical tasks. However, we decided to design our own escape room for two main reasons. First, developing a custom game allowed us to adapt the setup to our technical constraints, such as avoiding overly long sessions that could interfere with the robot's operation. Second, the custom design ensured better control over the types of tasks and the specific abilities they engage, allowing for a broader representation of cognitive and collaborative skills within a team. This controlled environment also increases replicability compared to traditional field-based escape rooms, which are harder to standardize and use for systematic evaluation of teamwork and problem-solving abilities.

Specifically, we incorporated four games, each targeting a distinct type of ability and skill set, while still drawing inspiration from conventional escape-room puzzles and hints (e.g., Mastermind-style code breaking and object-counting tasks). This approach ensures that the experience remains engaging and comparable to real-world escape-room logic, while maintaining full experimental control. The four games are as follows:

- Game 1: Linguistic skills are tested by identifying the intruder among seven anagrammed words.
- Game 2: Logical and mathematical reasoning are necessary for pattern recognition and numerical puzzles.
- Game 3: General knowledge of history and geography is required.
- Game 4: Logical reasoning abilities are assessed through a Mastermind-style code-breaking task, where players must find the correct code to obtain the key and exit the room.

A full description of each game and the corresponding solution logic is reported in the Appendix, together with the instructions provided to participants. An overview of the experimental setup is shown in Figure 1.

In our experiment, participants are grouped in teams of three with the main objective of escaping the room (that is, solving the 4 games) in 25 minutes (1,500 seconds). They are recruited online from a pool of more than 4,000 students in all departments at the University of Pisa through the ORSEE platform. Each participant receives an 11 Euro voucher as a participation reward to collect a T-shirt from the University store, regardless of their performance in the experiment. The group with the best performance every 25 participating teams will receive an additional voucher of 33 Euro for a hoodie.

Across experimental conditions, we vary both the team composition (all-human teams versus mixed teams of two humans and one artificial agent) and the embodiment of the artificial agent.

Fig. 1. Experimental setup of the virtual escape-room environment. The photo shows the arrangement of the main desk, the shared computer screen used by participants to view the puzzles and enter their answers, and the placement of the artificial agent (Abel). The monitor on the left corresponds to Abel's interface and was not accessible to participants. The locations of Hints 1–3 are indicated in the figure, while a detailed description of the content of each hint is provided in Appendix.

Table 1. Experimental Condition Overview

Condition	No embodiment	Low embodiment	High embodiment
Machine	Computer Box	Avatar	Abel
No Machine	Only Humans		

In particular, we use three different types of artificial agents (see Table 1 and Figure 2), each representing a different level of embodiment:

— *Computer Box*: A computer-based agent with no visual representation, serving as a baseline for the lowest level of embodiment. The Box functions without any anthropomorphic characteristics.
— *Avatar*: A virtual character that provides a moderate level of embodiment through its visual representation but lacks behavioral cues. The Avatar is designed to engage with human players using a digital persona that users can interpret or connect with, and it is very similar to the humanoid.[1]
— *Abel*: A hyper-realistic humanoid robot with a youthful appearance and non-specific gender, available at the Centro E. Piaggio at the University of Pisa [15]. ABEL represents the highest level of embodiment among the artificial agents, designed to closely mimic human behavior and appearance. However, in this setup, like the avatar, it does not incorporate gestures or facial expressions.[2]

Each participant wears an individual microphone to communicate with the artificial agent and with each other. This setup also ensured clean audio input for each participant, reducing background noise and overlapping speech, which are common issues in group interactions and can otherwise

[1]For more information and a demo, see https://github.com/phuselab/openFACS?tab=readme-ov-file and [18].
[2]For more information, see https://forelab.unipi.it/technologies/abel-new-generation-hyper-realistic-humanoid-robot.

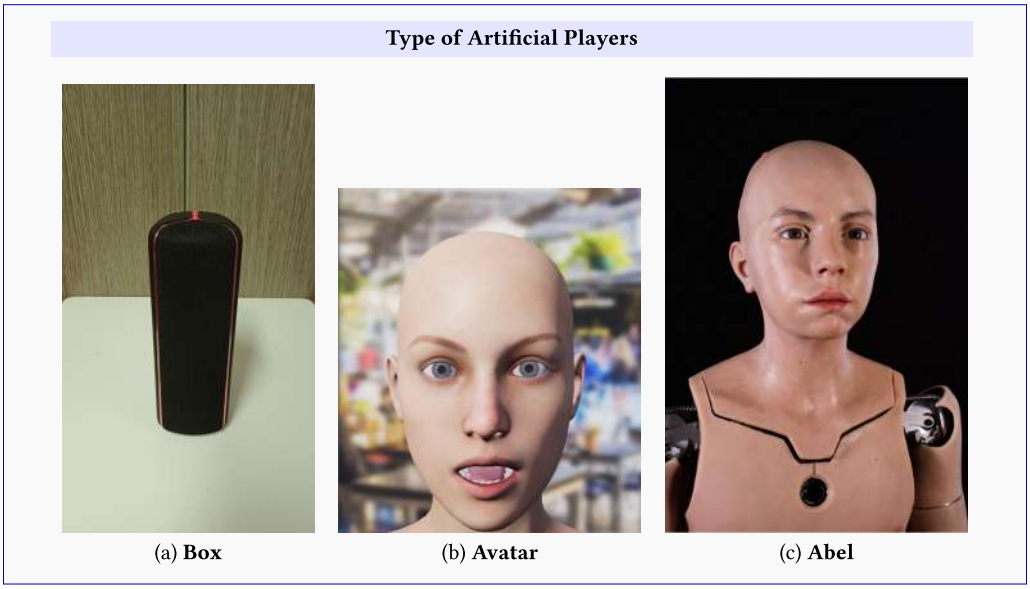


Fig. 2. Types of artificial players.

complicate the analysis of turn-taking and dialogue flow. Regardless of the embodiment level, interaction with the artificial agent is always verbal via an API to ChatGPT-4, with the agent responding in the same voice tone and providing answers both verbally and in written form. These responses are displayed on a shared screen in the room, where participants read the tasks, enter their answers, review the AI's responses, and track their own dialogue. The language model was initialized in each session with a fixed system prompt intended to promote comparable baseline behavior across all experimental conditions. The prompt instructed the artificial agent to adopt a polite and cooperative attitude and to act as a participant in the escape-room experiment. In particular, the prompt read (English translation): "*You are a participant in an escape-room session. Each session includes four games. You and the other players will interact together to solve each game. Act kindly and cooperatively with the other players and do not hesitate to ask questions if necessary.*" This fixed prompt ensured that the agent maintained a consistent level of social engagement and communication style, independent of its embodiment (Box, Avatar, or Abel).

In particular, when a player decides to interact with the artificial agent, the microphone used to initiate the conversation is isolated using JavaScript, and the audio is recorded. This audio is then sent to the backend, developed in Python, where it is converted into the .wav format. The audio file is subsequently transmitted via API to a server that uses Whisper, an advanced AI model for automatic speech-to-text transcription.[3]

After transcription, the text generated by Whisper is further processed: noise is filtered through a probability threshold that discards segments with a low probability of correctness, replacing them with "[Not Audible]." The resulting text is then sent via API to ChatGPT-4, which generates an appropriate response. This response is finally converted into audio using XTTS-v2, an advanced Text-to-Speech system.[4] The generated audio is then played back to the player. The large language

[3]Whisper is a deep neural network trained on a wide range of multilingual audio data, capable of accurately handling different accents, background noise, and variations in audio quality, converting speech to text with a high degree of accuracy.
[4]XTTS-v2 is designed to convert text into synthetic speech that sounds natural and fluid, closely resembling human voice quality.

models are integrated into our Otree code, borrowing some functionalities from the toolkit developed by [31].

To ensure fairness across all experimental conditions, none of the artificial agents (Box, Avatar, or Abel) had access to external information sources or online databases. Our API framework does not allow any form of external search or query, meaning that all agents operated only on the visual and textual information available within the experimental setup. This guarantees that any difference in performance reflects reasoning and interaction dynamics rather than unequal access to external knowledge.

3.2 Conversational Dynamics

To study conversational dynamics, we recorded and immediately transcribed the dialogue of each participant using individual microphones. This approach allows us to capture clear and distinct audio tracks for each participant, ensuring accurate transcription and analysis of the conversational flow.

While we provide a general overview of dialogue structure in all conditions, our analysis focuses on interactions that involve an artificial agent. This is both a methodological choice and a matter of scope. The all-human condition (three humans) naturally leads to different conversational dynamics, making direct comparisons with AI-mediated dialogues less meaningful. In human-only teams, participants engage in spontaneous, unstructured discussions, whereas in mixed conditions, interactions follow a structured question-and-answer format with the agent. Since our goal is to understand how human communication changes based on the AI's level of embodiment, and human-only interactions are inherently different, we focus on how AI presence and embodiment shape human interactions within mixed teams.

Specifically, to analyze conversational dynamics in mixed-team conditions, in the following, we utilize:

- Profiling-UD: It is a web-based tool conceived to stylistically profile multilingual texts by relying on the UD formalism [21], a *de facto* standard schema for morpho-syntactic and syntactic annotation of corpora, based on the dependency syntactic representation paradigm. It employs Computational Stylometry techniques based on Linguistic Profiling, a methodology that detects and quantifies differences and similarities across texts representing distinct language varieties by analyzing the distribution of numerous linguistic features [65]. As discussed in Section 5.1, by examining variations in the distribution of morpho-syntactic and syntactic properties computed by Profiling-UD within the collected conversations, we aim to explore whether different levels of embodiments of artificial agents influence the communication style adopted by humans.
- SBERT [57]: It is a modification of BERT [26] to produce sentence embeddings that can be easily compared to evaluate their similarity using cosine similarity, which ranges from −1 (dissimilar sentences) to 1 (identical sentences). By leveraging SBERT, we can quantify how closely participants' dialogues align in terms of meaning, with cosine similarity scores ranging from −1 (complete dissimilarity) to 1 (high similarity), where a score of 0 indicates no meaningful similarity. This scoring system allows us to compare semantic similarity across different experimental conditions, identifying patterns of agreement, disagreement, or topic continuity within conversations. In particular, by comparing SBERT scores across conditions, we can explore how the presence of an artificial agent influences the semantic flow of participants' dialogue with the agent.

By combining these analytical tools, we aim to develop a comprehensive understanding of the conversational dynamics at play when human teams collaborate with an artificial agent. These

insights will help us identify which key characteristics artificial agents need to have to enhance team communication and overall performance in non-routine, analytical tasks.

4 Experimental Evidence

In total, 174 students from all departments of the University of Pisa participated between April and June 2024 in the study. The experimental protocol was reviewed and approved by the Bioethics Committee of the University of Pisa (Review No. 25/2024), ensuring adherence to ethical standards throughout the research. The study involved 77 sessions across 4 experimental conditions: 3 humans only (20 sessions), 2 humans with a computer (*Box*) (16 sessions), 2 humans with *ABEL* (18 sessions), and 2 humans with an *Avatar* (23 sessions). We preregistered our analysis plan at https://osf.io/g2j7h.[5]

As our study has limited statistical power regarding performance measures and the number of conversations collected, we consider the evidence exploratory. Consequently, we acknowledge that the analysis presented in the following sections should be viewed as preliminary rather than conclusive. However, we believe that the collected dataset represents one of the very first resources for studying human–AI multimodal interaction, integrating audio and textual transcriptions of recorded conversations along with objective performance measures. Notably, this dataset will be made publicly available upon acceptance of the article at the associated OSF repository (https://osf.io/2v4ec/). In this respect, it provides a valuable initial resource for conducting multiple investigations into the dynamics of human–AI teaming in non-routine tasks.

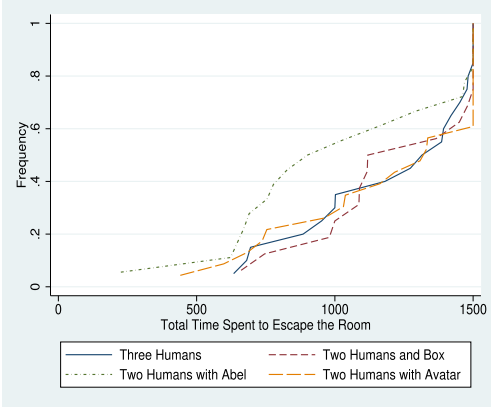
Table A1 in Appendix presents a summary of the statistics for the variables we collected at the end of the experiment through an exit questionnaire. These variables include individual characteristics such as a general risk-attitude item, adapted from [29], attitude towards technology (borrowed from [35]), escape game experience, and preference for the four games.[6] Additionally, we collected group characteristics by asking participants about their individual perceptions of aspects such as harmony and conflict during the interaction. A detailed list of the questions is provided at the end of the article. Half of our sample consists of students aged between 21 and 50, 47% are female, and 43% come from a STEM field. The average risk propensity is approximately 2.2 out of 4, while the average self-reported measure of regular AI use is about 5 out of 7. Preferences for each of the four games are equally distributed, and 35% had prior experience with an escape room game. The majority of groups reported a high level of harmony, with an average score of 6 out of 7.

4.1 Results: Team Performance

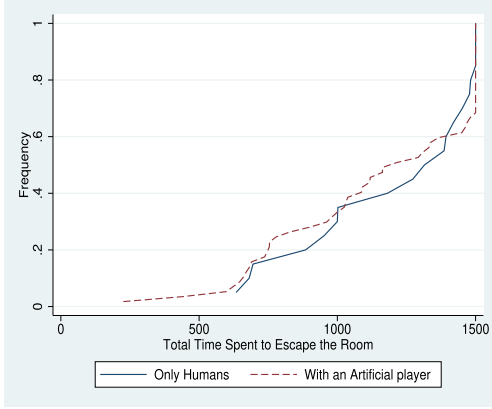
The main indicator of team performance is the total time spent (in seconds) to exit the room and complete each subgame. Figure 3 presents the **Empirical Cumulative Distribution (ECDF)** of overall finishing times across conditions, with the x-axis representing completion time (in seconds) and the y-axis showing the cumulative proportion of teams that completed the task within that time. The ECDF curves provide a comprehensive picture of performance distribution across all teams under the 1,500-second time limit. The cumulative share of teams at the very end of the distribution represents those that successfully completed the game, while the complement corresponds to those who failed. Figure 4 is analogous, but the ECDF is presented separately for each subgame.

[5] Three sessions (one all-human and two with artificial players) were lost due to technical issues such as microphone malfunctions or software errors and were therefore excluded from the analyses. One additional session was retained despite missing end-time data for each subgame, as all other information was complete.

[6] As in our previous work [14], we collected a single post-experimental human-likeness item, which was not statistically significant in the current experiment. We do not consider this measure a validated embodiment or anthropomorphism scale, especially for modern language-model-based agents, and we return to this measurement limitation in the Discussion.

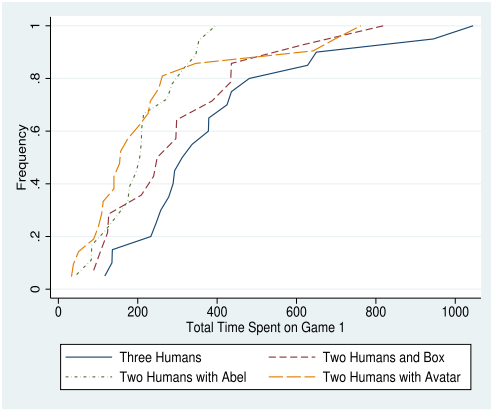


(a) Human-only teams

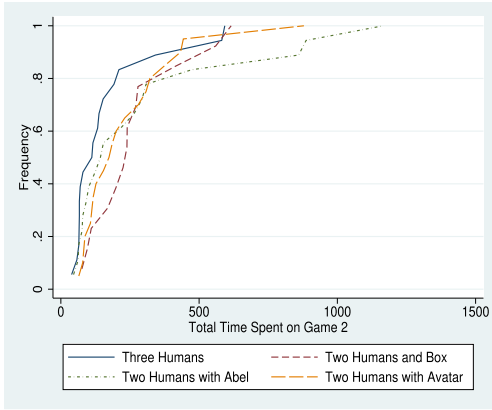


(b) Team w/o an artificial player

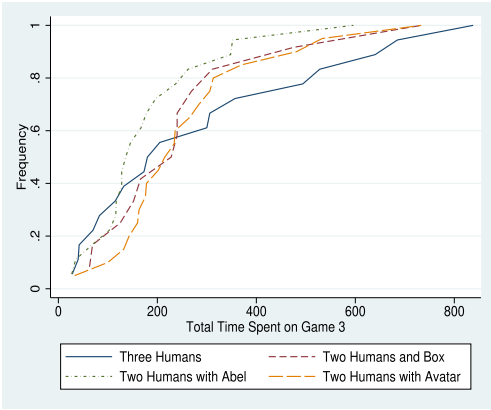
Fig. 3. Total time spent to escape the room by condition.



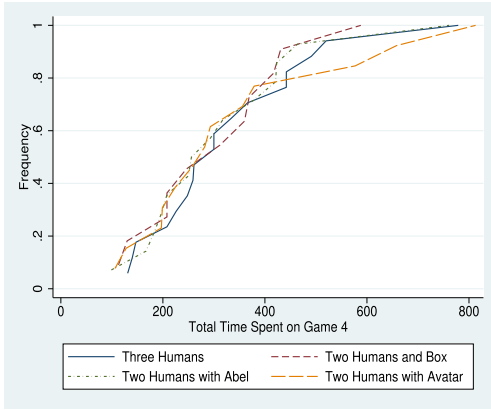
(a) Game 1



(b) Game 2



(c) Game 3



(d) Game 4

Fig. 4. Time spent on each game by condition.

Two important aspects emerge from Figure 3, panel (a). At the beginning, the ECDFs for teams with higher levels of embodiment, Abel and the Avatar, tend to lie above the others. This indicates that, for each time threshold (e.g., 700 seconds), the proportion of teams that managed to escape the room was higher when an embodied artificial teammate was present. However, toward the end of the distribution (around 1,500 seconds), the curves for the human-only teams become steeper. Focusing on this final segment reveals that the cumulative share of human-only teams that successfully escaped the room in the end (about 85%) is higher than that of teams with an artificial player (Box about 69%, Avatar about 57%, and Abel about 78%). These differences emerge also in panel (b), where results across all conditions with artificial players are pooled together (Humans 85% vs. with an artificial player 67%). The curve for teams with an artificial player is thus wider, starting earlier and extending further along the time axis, reflecting greater variability in completion times when an artificial player was included, with a small number of teams performing exceptionally well. Furthermore, we report the empirical distributions of finishing times for each of the four subgames by condition in Figure 4. The most pronounced differences emerge in Game 1, where the presence of an artificial player appears to substantially enhance team performance, whereas these differences tend to diminish progressively across subsequent games.

At first glance, this pattern might seem contradictory: human-only teams complete the task more frequently within the 1,500-second limit, yet mixed teams display both very fast and very slow completion times. Across embodiment levels, a non-monotonic pattern also emerges: the intermediate Avatar condition shows the lowest overall success despite similar early-stage performance and the presence of some teams that performed exceptionally well. These results, which are consistent with the uneven impact of generative AI observed in non-routine entrepreneurial tasks (see [54]), suggest that embodiment affects not only average outcomes but also their variability, likely depending on how effectively each team integrates and coordinates with the agent, a point we return to in the Discussion. Concerning differences across tasks, the important difference in Game 1 is likely related to the nature of the task, which requires linguistic reasoning, an area that falls naturally within the capabilities of generative AI systems such as ChatGPT (as stated in Section 3.1, our API framework does not allow any form of external search or query). At the same time, the first game includes all participating teams, since no selection has yet occurred. This means that the observed advantage may also reflect the AI's ability to help teams overcome initial impasses, particularly in identifying the "intruder," and thereby avoid becoming stuck early in the session. We cannot fully disentangle these factors, but both the linguistic component of the task and the absence of prior selection plausibly contribute to the stronger AI effect observed in this stage.

To support our analysis, Table 2 presents statistical tests comparing the time distributions of humans and artificial agents across the four subgames, the upper part considering all teams together, while the bottom part considering only teams that successfully completed all four games and escaped the room. Specifically, the **Anderson-Darling (AD)** and **Kolmogorov-Smirnov (KS)** tests assess whether finishing times differ significantly across team compositions. The KS test measures the maximum difference between cumulative distribution functions, providing a global comparison, while the AD test is more sensitive to differences in the tails, making it particularly useful for detecting extreme performances (e.g., exceptionally fast or slow escape times).[7]

Across all games, the AD (Stat $= 53.21, p = 0.60$) and KS (Stat $= 0.14, p = 0.82$) tests show no significant differences between the time distributions of human and artificial-agent teams, suggesting overall similarity when considering all teams. The mean times (Humans = 1,211.45, Artificial = 1,153.02) further confirm this comparability. However, the distribution tends to differ

[7] Additionally, the AD test requires less data than the KS test to achieve sufficient statistical power. See e.g. [7, 33].

Table 2. Comparison of Completion-time Distributions across Team Types

Variable	AD Statistic	AD p-value	KS Statistic	KS p -value	Mean Humans	Mean Artificial	AD BH p-value	KS BH p-value
All Teams								
Total time spent	53.208	0.616	0.141	0.820	1211.45	1153.02	0.769	0.936
Game 1 time spent	76.219	0.004	0.454	0.003	400.05	248.92	0.020	0.015
Game 2 time spent	70.608	0.032	0.330	0.085	171.28	268.04	0.080	0.212
Game 3 time spent	66.101	0.582	0.204	0.549	289.78	233.40	0.769	0.914
Game 4 time spent	52.493	0.841	0.136	0.936	329.06	322.84	0.841	0.936
Successful teams only								
Total time spent	55.869	0.046	0.325	0.129	1161.47	979.53	0.116	0.215
Game 1 time spent	56.916	0.002	0.455	0.009	373.18	213.76	0.007	0.046
Game 2 time spent	55.010	0.146	0.333	0.107	170.06	223.21	0.244	0.215
Game 3 time spent	53.483	0.527	0.248	0.388	289.18	219.71	0.659	0.485
Game 4 time spent	52.493	0.850	0.136	0.936	329.06	322.84	0.850	0.936

This table compares completion-time distributions between human-only teams and teams with one artificial teammate. The top panel reports results for all teams; the bottom panel restricts the sample to successful teams only. For each variable, we report AD and KS two-sample tests and their raw p-values. The final columns show BH adjusted p-values. Small differences may arise from rounding. AD, Anderson-Darling; BH, Benjamini-Hochberg; KS, Kolmogorov-Smirnov.

significantly when considering only successful teams, as highlighted by the AD (Stat = 55.87, p = 0.046) that tend to capture differences in the tails of the distribution.

In addition, Game 1 stands out as the only task showing a robust difference in completion-time distributions between human-only and mixed teams. Both the AD test (p = 0.004 for all teams, and 0.020 for successful teams) and the KS test (p = 0.003 for all teams, and 0.009 for successful teams) indicate significant contrasts, with humans taking longer on average (400.05 vs. 248.92 for all teams; 373.18 vs. 213.76 for successful teams). This suggests that artificial agents facilitated faster completion of the first puzzle.

As shown in the last two columns of Table 2, Game 1 remains significant after applying the **Benjamini-Hochberg (BH)** correction for multiple testing. By contrast, neither total completion time nor the remaining games (Games 2–4) show significant differences once corrected, indicating similar performance across team types for the later stages of the task sequence. Differences in success rates are small and not statistically robust. Proportion tests have limited power with about 20 teams per condition, so we do not report them and focus instead on the more informative distributional analysis.

To sum up, teams including an artificial agent displayed greater variability in performance—some completing the tasks very rapidly, others struggling to finish within the time limit, whereas human-only teams tended to perform more consistently. The most pronounced difference appears in Game 1, where the presence of an artificial teammate significantly reduced completion times. In subsequent games, as only the more proficient teams advanced, this advantage diminished, suggesting that selection effects became increasingly important. Taken together, these results indicate that the positive impact of artificial teammates is strongest in the early stages, when teams face greater uncertainty and coordination challenges.

4.1.1 Number of Errors. Average number of errors as an additional measure of performance, we now focus on the number of errors teams committed before providing the correct answer. Because not all teams completed the same number of games, we compute for each team the average number of errors per completed game. This normalization ensures that performance is comparable across teams with different levels of task completion. Table 3 reports the differences in this average across

Table 3. T-Test Results for Errors across Games by Experiment Type

	Mean 1	N1	Mean 2	N2	Diff	p-Value
Average errors						
Humans vs. Box	7.30	20	4.68	14	2.62	0.43
Humans vs. Abel	7.30	20	2.44	18	4.86	0.08
Humans vs. Avatar	7.30	20	4.33	21	2.97	0.26
Box vs. Abel	4.68	14	2.44	18	2.24	0.13
Box vs. Avatar	4.68	14	4.33	21	0.35	0.84
Abel vs. Avatar	2.44	18	4.33	21	−1.89	0.11
Only humans	7.30	20	3.78	53	3.52	0.05
Game 1 errors						
Humans vs. Box	7.85	20	9.12	16	−1.28	0.83
Humans vs. Abel	7.85	20	3.22	18	4.63	0.24
Humans vs. Avatar	7.85	20	13.22	23	−5.37	0.54
Box vs. Abel	9.12	16	3.22	18	5.90	0.22
Box vs. Avatar	9.12	16	13.22	23	−4.09	0.68
Abel vs. Avatar	3.22	18	13.22	23	−10.00	0.25
Only humans	7.85	20	8.91	57	−1.06	0.86
Game 2 errors						
Humans vs. Box	2.70	20	3.87	15	−1.17	0.49
Humans vs. Abel	2.70	20	2.72	18	−0.02	0.99
Humans vs. Avatar	2.70	20	2.33	21	0.37	0.79
Box vs. Abel	3.87	15	2.72	18	1.14	0.48
Box vs. Avatar	3.87	15	2.33	21	1.53	0.29
Abel vs. Avatar	2.72	18	2.33	21	0.39	0.77
Only humans	2.70	20	2.89	54	−0.19	0.87
Game 3 errors						
Humans vs. Box	7.50	18	2.92	13	4.58	0.10
Humans vs. Abel	7.50	18	1.89	18	5.61	0.01
Humans vs. Avatar	7.50	18	5.05	20	2.45	0.34
Box vs. Abel	2.92	13	1.89	18	1.03	0.36
Box vs. Avatar	2.92	13	5.05	20	−2.13	0.30
Abel vs. Avatar	1.89	18	5.05	20	−3.16	0.06
Only humans	7.50	18	3.39	51	4.11	0.02
Game 4 errors						
Humans vs. Box	2.72	18	1.36	11	1.36	0.25
Humans vs. Abel	2.72	18	1.33	15	1.39	0.15
Humans vs. Avatar	2.72	18	3.18	17	−0.45	0.75
Box vs. Abel	1.36	11	1.33	15	0.03	0.97
Box vs. Avatar	1.36	11	3.18	17	−1.81	0.27
Abel vs. Avatar	1.33	15	3.18	17	−1.84	0.18
only humans	2.72	18	2.07	43	0.65	0.51

"Mean 1" and "Mean 2" represent the average number of errors made by teams in Conditions 1 (e.g., Human) and 2 (e.g., Box), respectively. "N1" and "N2" indicate the number of observations for each condition. "Diff" refers to the difference in mean errors between the two conditions, while "p-value" represents the statistical significance of this difference based on a t-student test.

conditions, as well as the same measure computed separately for each subgame. To assess these differences, we conducted two-sample Welch t-tests comparing the mean number of errors between pairs of conditions.[8]

The results indicate differences in team performance depending on their composition. On average, human-only teams made significantly more errors than teams that included an artificial agent (7.30 vs. 5.78). This difference is both economically (3.52) and statistically significant (p-value = 0.05). This finding is robust to a permutation test with 500 permutations, and remains consistent when using the total number of errors instead of the average, suggesting that the presence of an artificial agent—regardless of its embodiment level, tends to reduce the number of errors made by the team.

When comparing human-only teams to each specific type of artificial agent (Box, Avatar, and Abel), or within each subgame, the same qualitative pattern holds: human teams typically make more errors, but most differences are not statistically significant individually (p-values 0.10–0.18), reflecting limited power in subgroup comparisons. The exception is Game 3. Here, human-only teams made substantially more errors (difference = 3.39, p-value = 0.02), and there is an advantage of playing with Abel (difference = 5.61, p-value = 0.01). These results also persist under the 500-permutation test.

Finally, we find suggestive evidence of a non-linear pattern: the highest error rates occur at intermediate embodiment levels, consistent with the broader evidence of a "valley" in mixed-team coordination. This suggests that while artificial agents generally help reduce errors, the specific type of artificial teammate can meaningfully shape team performance.[9]

Summary of Performance Results. Team performance exhibits a counterintuitive pattern. Mixed human–AI teams often start by solving puzzles faster, especially with more embodied agents, yet human-only teams are ultimately more likely to complete all tasks within the 1,500-second limit. This contrast does not reflect systematically worse performance by mixed teams, but rather greater variability in how effectively human–AI groups coordinate. Human-only teams, by comparison, perform more consistently.

Three main results emerge:

(1) *Overall Completion Times* (ECDF analysis). Mixed teams show earlier and sometimes very fast completion times, but they also experience more extreme delays. Human-only teams, despite starting slower, achieve a higher final success rate (about 85% vs. about 67% for mixed teams). In other words, mixed teams exhibit much higher variability: some perform extremely well while others struggle, which is what makes the overall pattern appear counterintuitive.
(2) *Game-Level Differences*. The strongest contrast appears in *Game 1*, a linguistically demanding task where teams with an artificial agent finish substantially faster, a difference that remains significant after applying the BH correction. In later games (2–4), these differences diminish as selection effects accumulate, as only the most proficient teams advance.
(3) *Errors.* Human-only teams commit more errors on average, a result robust to permutation tests. The largest effect occurs in *Game 3*. Results also suggest a non-linear embodiment pattern, with intermediate embodiment associated with the highest error rates.

[8]The t-test statistic for comparing two independent samples is $t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$, where $\bar{x}$ and $\bar{y}$ are the sample means, s_x^2 and s_y^2 are the sample variances, and n_x and n_y are the sample sizes. The test follows a Student's t distribution with degrees of freedom given by Welch's approximation.

[9]An additional analysis (available upon request) categorizes teams into "Success" and "Failure" based on game completion. The results reinforce that human-only teams make more errors than those with an artificial agent, even when successful. Moreover, findings suggest a non-linear relationship between error count and agent embodiment, though this interpretation is constrained by the study's statistical power.

Table 4. Performance Variables and Individual and Group Characteristics

	Model 1			Model 2			Model 3			Model 4		
	Success	Time (a)	Errors	Success	Time (b)	Errors	Success	Time (c)	Errors	Success	Time (d)	Errors
Box	0.688***	1224.7***	16.06***	0.487***	1555.2***	34.90***	0.733***	1232.8***	9.781	0.589***	1300.7***	27.11***
	(0.0768)	(58.67)	(3.971)	(0.160)	(117.0)	(8.221)	(0.171)	(131.6)	(8.875)	(0.0946)	(72.72)	(4.760)
Abel	0.778***	1029.4***	8.944**	0.571***	1378.3***	28.08***	0.799***	1037.2***	3.486	0.638***	1133.4***	23.77***
	(0.0724)	(55.32)	(3.744)	(0.159)	(116.7)	(8.200)	(0.172)	(132.3)	(8.923)	(0.104)	(80.13)	(5.245)
Avatar	0.565***	1199.9***	22.09***	0.370**	1537.8***	40.02***	0.590***	1204.7***	16.71*	0.451***	1286.2***	34.49***
	(0.0640)	(48.94)	(3.312)	(0.148)	(108.5)	(7.626)	(0.165)	(126.9)	(8.556)	(0.0895)	(68.83)	(4.506)
Only humans	0.850***	1211.5***	19.75***	0.646***	1557.4***	38.90***	0.884***	1223.8***	13.76	0.658***	1340.6***	37.02***
	(0.0561)	(42.85)	(2.900)	(0.154)	(112.5)	(7.909)	(0.164)	(126.3)	(8.515)	(0.107)	(82.24)	(5.383)
Group: friends				−0.0105	−4.336	0.520						
				(0.0164)	(12.03)	(0.846)						
Group: connection				0.0481*	−66.54***	−3.409***						
				(0.0245)	(17.91)	(1.259)						
Group: conflicts				−0.0261	33.31	−0.917						
				(0.0322)	(23.59)	(1.658)						
Regular AI use							−0.0275	10.56	0.761			
							(0.0185)	(14.24)	(0.960)			
Tech curiosity							−0.0163	−7.208	1.136			
							(0.0197)	(15.17)	(1.023)			
Tech propensity							0.0276	−5.910	−0.289			
							(0.0268)	(20.67)	(1.394)			
Escape experience										−0.0538	−7.422	−4.981
										(0.0691)	(53.17)	(3.480)
Game entropy (norm.)										0.402**	−247.5*	−31.02***
										(0.178)	(136.9)	(8.960)
Observations	174	174	174	174	174	174	174	174	174	174	174	174

Standard errors are reported in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Pure differences in overall success rates across team types are not always statistically significant, but given the sample size, the minimum detectable effect exceeds the observed gap, indicating limited statistical power rather than true equivalence.

4.1.2 Controlling for Individual Characteristics. To assess whether our main results are robust, we estimate a regression model considering our three key performance metrics: success rate (whether the team escaped), time taken to escape, and the number of errors, and controlling for individual characteristics, such as gender, prior experience with escape rooms, and familiarity with technological tools.

Table 4, column (a), presents the estimated effects of AI embodiments (Box, Avatar, Abel) along with those for the all-human team. This setting essentially replicates the results from Tables 2 and 3, but with dummy variables for each condition. In column (b), we introduce group-level controls related to the role of team dynamics in performance outcomes. Specifically, we account for whether team members were friends prior to the experiment (Group Friend) and subjective perceptions of team connection and conflicts.[10] These factors may be relevant for collaboration. Although these measures are only exploratory and should be interpreted with caution, a stronger sense of connection is associated with a higher probability of success, shorter completion times, and fewer errors, while preserving the effect of each level of embodiment.

In column (c), we further extend our model by incorporating individual attitudes toward technology. We include controls for prior AI experience (AI User), technological curiosity, and technological propensity to assess whether familiarity with AI and comfort with digital tools are correlated with how participants interact with artificial teammates. None of these variables appears significant.

Finally, column (d) controls for individual preferences for each specific game and for prior escape-room experience, both of which could be related to performance. We construct an

[10] The variable group harmony is included in unreported regressions. The results remain identical; however, we do not present this specification because the variable is missing for a small subset of participants.

Table 5. Average Messages Exchanged by Experiment Type (per Human Participant)

Experiment conditions	Game 1	Game 2	Game 3	Game 4	Total
Humans	3.95	1.52	0.98	1.12	7.57
Box	1.47	0.69	0.91	0.25	3.31
Avatar	0.96	0.30	1.20	0.61	3.07
Abel	0.58	0.53	0.81	0.39	2.31
Overall	1.75	0.75	0.99	0.61	4.11

entropy-based indicator of game preference, derived from each participant's stated preferences across the available games. An increase in game-preference entropy by 1 (i.e., moving from zero to full entropy) is associated with a 0.402 increase in the probability of successfully escaping from the room, a reduction in completion time of about 247 seconds, and a decrease in the total number of errors by about 31 on average.

To sum up, controlling for individual, group, and game characteristics does not alter our main conclusions. At the same time, the evidence based on these additional controls remains exploratory and correlational. In particular, the most suggestive pattern is that teams exhibiting fewer interpersonal conflicts and greater variability in members' game preferences tend to perform better. We return to this point in the Discussion.

5 Results: Conversation Dynamics

After analyzing the performance of both human-only and hybrid human–AI teams in escaping the room, we now turn to examine the dynamics of conversations where humans interacted with their teammates during the administered games.

As a first step, we provide a general overview of the number of messages exchanged by each human participant in each session and across all experimental conditions. This overview allows for a comparative analysis to investigate whether changes in experimental conditions or game type influence communication frequency. Specifically, we aim to determine whether the presence of an artificial teammate, along with different levels of embodiment, and the type of cognitive skills required to solve the game, affect participants' propensity to communicate. The corresponding statistics are reported in Table 5.

As shown in the table, humans tend to interact more frequently with each other compared to when they interact with artificial agents. Specifically, during our experiments, the total average number of messages exchanged among humans, across all games, is 7.57, significantly higher than the average number of messages exchanged with Box, Avatar, and Abel. Interestingly, a ranking can be observed across these three conditions, which appears to correlate with the agents' levels of embodiment. Humans exchanged a greater average number of messages with the least embodied agent, Box, and the fewest with the most highly embodied agent, Abel.

As mentioned above (see Section 3.2), the human-only condition is inherently different in dialogue structure, as participants engage in conversations with each other. In contrast, in the conditions with artificial agents, participants interact in a question-and-answer format with the artificial agent. Therefore, the communication in the human-only setting is not directly comparable to the other three, i.e., Box, Avatar, and Abel. For this reason, in what follows, we focus our comparative analysis only on the latter conditions. Specifically, Section 5.1 reports and discusses the results of the analysis of the communication style adopted by humans when interacting with the three artificial agents, while in Section 5.2, we discuss the outcomes of an analysis focused on the content of the messages exchanged by humans across the three experimental conditions.

Table 6. Linguistic Features Extracted by Profiling-UD Categorized into Nine Groups of Linguistic Phenomena

Raw text properties
Number of sentences
Average sentence length (in terms of words per sentence, including punctuation marks)
Average word length (in terms of characters per word)
Vocabulary richness
Type/Token Ratio (TTR) for words and lemmas
Morphosyntactic information
Distribution of the 17 core part-of-speech categories defined in the Universal tagset
Lexical density
Verbal inflectional morphology
Distribution, for each lexical verb and auxiliary, of the following subset of inflectional features: Verb form, Mood, Tense, Number, and Person
Verbal predicate structure
Distribution of verbal heads and verbal roots
Average verb arity, i.e., number of dependency links governed by a verbal head) and distribution of verbs by arity
Global and local syntactic tree structures
Average depth of the whole syntactic tree
Average length of dependency links and of the longest link
Number of embedded complement chains governed by a nominal head
Average depth of embedded complement chains and distribution of chains by depth
Average clause length
Relative order of elements
Distribution of pre- and post-verbal subjects and direct objects
Syntactic relations
Distribution of the 37 dependency relations defined in the UD Universal tagset
Use of subordination
Distribution of subordinate and principal clauses
Average depth of embedded subordinate clauses and distribution of subordinate “chains” by depth
Distribution of subordinate clauses preceding and following the principal clause

5.1 Analysis of Communication Style

As previously mentioned, for the analysis of the communication style of the messages, we relied on Profiling-UD [9], a web-based Computational Stylometry tool that operates through a two-stage process: linguistic annotation and linguistic profiling. The first stage, linguistic annotation, is automatically handled by UDPipe [61], a state-of-the-art pipeline available for nearly all languages in the UD initiative. UDPipe performs basic pre-processing tasks, such as sentence splitting, tokenization, Part-of-Speech tagging, lemmatization, and dependency parsing. In the second stage, approximately 130 features representing the linguistic structure of the text are extracted from the output of the various levels of linguistic annotation. These features capture a wide range of linguistic phenomena, proven effective in various contexts focused on the style of a text. The features are categorized into nine groups, each corresponding to different linguistic phenomena. As shown in Table 6, they model aspects ranging from raw text properties to morpho-syntactic information and the inflectional properties of verbs. Additionally, more complex aspects of sentence structure are considered, including both global and local properties of parsed trees and specific subtrees, such as the order of subjects and objects relative to the verb, the distribution of UD syntactic relations, and features related to subordination and the structure of verbal predicates.

Our linguistic profiling analysis focused exclusively on the transcriptions of what the two human participants in each team communicated to the artificial agent, across the conditions considered. Thus, we excluded the responses of the artificial agents, as our primary research objective is to investigate whether the agent’s embodiment influences the communicative style adopted by humans. In contrast, we were not concerned with analyzing the communication style of the agents’ responses, regardless of their embodiment. Additionally, we averaged the results across games, as our focus was on the human communication style when interacting with artificial agents, rather than on game-specific variations.

Table 7. Average Values of Raw Text Features Extracted from Messages Exchanged between Humans and the Three Artificial Agents

Raw text feature	Box	Avatar	Abel
Number of sentences	1.05	1.21	1.55
Number of words	10.54	13.48	18.70

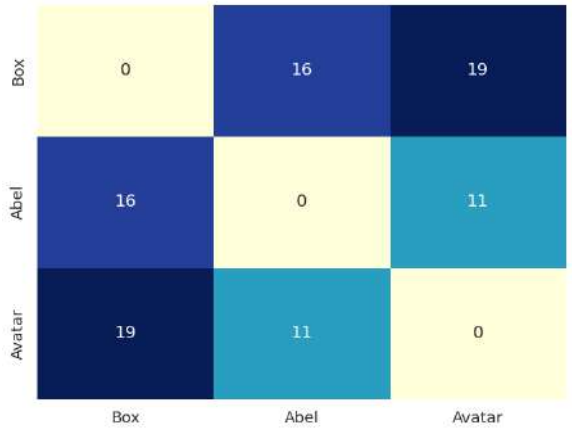


Fig. 5. The number of statistically different features extracted by Profiling-UD across three experimental conditions, excluding the human-only condition.

Table 7 presents the results of the initial analysis, which focuses exclusively on the raw text properties of the transcribed messages. It reveals that humans tend to produce a higher number of sentences and words when interacting with Abel compared to Box and Avatar. This finding introduces an interesting aspect of communication style based on raw text characteristics: although humans exchange the fewest number of messages with Abel on average (see Table 5), these messages are longer in terms of sentences and words.

Then, we conducted a statistical analysis to determine whether significant differences exist in the distribution of the linguistic features extracted by Profiling-UD from the humans' utterances across the three conditions. We used the Mann-Whitney U rank test for independent samples, which revealed that 46 features varied significantly ($p < 0.05$) across the following paired comparisons where two humans are involved: Box vs. Avatar, Box vs. Abel, and Avatar vs. Abel. This quantitative result seems to confirm our initial intuition, suggesting that humans tend to alter certain aspects of their communication style depending on the level of embodiment of the artificial agents with which they interact. The confusion matrix showing the number of linguistic features, regardless of their type, is presented in Figure 5. Interestingly, the majority of features vary when comparing the two most distinct embodiments, i.e., Box vs. Avatar and Box vs. Abel. The statistical significance test reveals that 19 linguistic features differ when humans interact with Box rather than Avatar, and 16 features differ when interacting with Box rather than Abel. In contrast, when humans interact with Avatar rather than Abel (i.e., Avatar vs. Abel), fewer characteristics of their communication style change, specifically 11 features. This suggests that human utterances tend to exhibit more similarities when the levels of embodiment of the artificial agents are higher and more similar to each other, such as Avatar and Abel.

Table 8. Linguistic Features That Are Statistically ($p < 0.05$) Different between Abel and Box Conditions

Group	Feature	Box	Abel	r
Raw Text	Number of sentences*	1.05 ± 0.21	1.55 ± 1.7	−0.16
Morphosyntax	% auxiliaries	9.13 ± 7.83	6.75 ± 7.77	0.19
	% interjections	2.39 ± 9.81	0.15 ± 0.99	0.07
	% proper nouns	5.81 ± 11.99	8.35 ± 11.99	−0.17
Inflectional morphology	% auxiliaries:present tense	65.09 ± 46.67	52.69 ± 48.9	0.15
	% verbs:gerundive mood	4.01 ± 17.89	0.0 ± 0.0	0.06
	% verbs:imperative mood*	0.94 ± 9.67	8.84 ± 27.91	−0.09
	% verbs:2nd-person-singular	7.55 ± 26.42	14.83 ± 33.23	−0.12
Vocabulary Richness	TTR:lemma-100 words	0.0 ± 0.0	0.01 ± 0.08	−0.04
	TTR:form-100 words	0.0 ± 0.0	0.02 ± 0.09	−0.04
Syntactic Relations	% predeterminers	0.13 ± 1.38	1.02 ± 4.15	−0.05
Order of elements	% post-verbal direct objects	17.06 ± 35.8	27.98 ± 42.64	−0.13
Tree Structures	% complement chains with depth 1	26.42 ± 44.09	41.39 ± 48.73	−0.15
	Avg depth of complement chains	0.33 ± 0.56	0.5 ± 0.58	−0.17
	Number of complement chains	0.3 ± 0.48	0.55 ± 0.81	−0.17

Mean values and standard deviations (±) are reported. Features in each group are ordered by absolute decreasing rank-biserial correlation value (r).
*Marks highly statistically significant features ($p < 0.01$).

We further focused our analysis on the linguistic features that vary significantly in each paired comparison. To identify the features with the greatest differences, we computed their rank-biserial correlation score r [68], which ranges from −1 to +1.[11] The absolute value of r indicates the magnitude of the distribution difference between the two experimental conditions, while the sign (positive or negative) indicates the direction of the difference in feature values between the two conditions. Tables 8–10 present the set of features with distributions that are statistically different in each of the three comparisons, ordered by decreasing absolute r score, along with their distribution values and standard deviations for each of the two compared conditions. In addition, to combine a primarily descriptive perspective with statistically robust evidence, we applied a **False Discovery Rate (FDR)** correction using the BH procedure. The correction was computed separately for the comparisons between each experimental setting (i.e., each paired condition) and for each of the nine groups of linguistic features considered in the analysis. This approach allowed us to identify the features that are most distinctive across settings, those that exhibit the strongest variation, while maintaining the descriptive focus of the study, which is primarily aimed at detecting meaningful trends in human conversational behavior.

In line with the results obtained from the inspection of the raw textual features, we hypothesized that the level of embodiment of the artificial agent influences the overall linguistic structure of team communication. Specifically, we expected that higher embodiment would prompt human participants to adopt a communication style characterized by linguistic features suggesting that they treat the artificial agent less as a tool to query for correct answers and more as a teammate to engage in interactive exchange. Accordingly, we hypothesized that the features varying most significantly between utterances produced by participants interacting with Abel (the most embodied agent) and those interacting with either Avatar or Box would reflect a more "natural" conversational style within the problem-solving framework of the escape room.

As a general remark, we observe that, regardless of the condition, the standard deviation values are relatively high. This is expected, as the conversations involve different individuals, each potentially characterized by distinct personal communication styles. Despite these individual differences, our

[11]Note that r is a score capturing the effect size of the Mann-Whitney U rank test, meaning that it provides a standardized way to interpret the practical significance of the results, complementing the p-value by quantifying the strength of the observed difference.

Table 9. Linguistic Features That Are Statistically ($p < 0.05$) Different between Avatar and Box Conditions

Group	Feature	Box	Avatar	r
Morphosyntax	% adpositions	7.14 ± 7.33	8.99 ± 7.35	0.15
	% adjective	3.88 ± 6.83	5.21 ± 6.67	0.14
	% adverbs	4.97 ± 14.27	1.96 ± 4.75	−0.12
	% punctuation marks	15.29 ± 12.49	12.25 ± 8.39	−0.15
Inflectional morphology	% verbs:2nd-person-singular*	7.55 ± 26.42	20.02 ± 38.48	0.15
	% verbs:finite forms	29.09 ± 40.8	41.74 ± 45.25	0.14
	% verbs:imperative mood*	0.94 ± 9.67	9.57 ± 29.12	0.09
	% auxiliaries:1st-person-plural	9.12 ± 27.2	3.43 ± 17.25	−0.07
Syntactic relations	% direct objects	3.55 ± 7.06	5.1 ± 6.63	0.16
	% determiners	13.76 ± 9.68	16.27 ± 7.71	0.16
	% case-marking	6.71 ± 7.35	8.58 ± 7.32	0.15
	% punctuation	15.29 ± 12.49	12.25 ± 8.39	−0.15
Order of elements	% post-verbal direct objects	17.06 ± 35.8	27.67 ± 41.77	0.13
Tree structures	Number of complement chains	0.3 ± 0.48	0.49 ± 0.6	0.16
	Average clause length	6.02 ± 3.59	6.82 ± 3.31	0.15
	Avg depth of complement chains	0.33 ± 0.56	0.49 ± 0.61	0.15
	% complement chains with depth 1	26.42 ± 44.09	39.36 ± 48.67	0.13
Subordination	% principal clauses	63.84 ± 38.71	74.44 ± 36.0	0.15
Predicate structure	% verbs with arity 2	23.89 ± 40.56	33.64 ± 42.69	0.13

Mean values and standard deviations (±) are reported. Features in each group are ordered by absolute decreasing rank-biserial correlation value (r).
*Marks highly statistically significant features ($p < 0.01$).

Table 10. Linguistic Features That Are Statistically ($p < 0.05$) Different between Abel and Avatar Conditions

Group	Feature	Abel	Avatar	r
Raw text	Number of sentences	1.55 ± 1.7	1.21 ± 0.86	−0.11
Morphosyntax	% determiners*	13.02 ± 9.5	16.46 ± 7.81	0.25
	% proper nouns*	8.35 ± 11.99	5.6 ± 12.66	−0.19
	Lexical density	0.5 ± 0.17	0.44 ± 0.14	−0.19
	% punctuation marks*	15.76 ± 10.58	12.25 ± 8.39	−0.24
Syntactic relations	% determiners*	11.98 ± 7.81	16.27 ± 7.71	0.3
	% predeterminers	1.02 ± 4.15	0.03 ± 0.4	−0.05
	% conjuncts	3.48 ± 7.46	1.49 ± 4.51	−0.11
	% punctuation*	15.78 ± 10.55	12.25 ± 8.39	−0.24
Predicate structure	% verb with arity 2	20.16 ± 34.88	33.64 ± 42.69	0.14
	% verbal roots	76.27 ± 38.98	86.54 ± 32.0	0.13

Mean values and standard deviations (±) are reported. Features in each group are ordered by absolute decreasing rank-biserial correlation value (r).
*Marks highly statistically significant features ($p < 0.01$).

focus is on the average feature values. Starting with the analysis of conditions where the two humans converse with artificial agents with the most distinct levels of embodiment (Abel vs. Box), Table 8 shows that humans tend to use more proper nouns, second-person singular verbs, present tense, and imperative forms when interacting with Abel rather than Box. Upon inspecting the conversations, we find that people tend to address Abel by name and in the second person, as if it were a human (e.g., "Abel, ci aiuti con l'indovinello?," transl. *Abel, can you help us with the riddle?*; "Ciao Abel, qual è la capitale della Cipro Turca?," transl. *Hi Abel, what is the capital of Turkish Cyprus?*; "Abel stai zitto ti prego," transl. *Abel please shut up*). Additionally, people tend to engage in longer conversations with Abel compared to Box, as indicated by the higher number of sentences per conversation. Consider for example the following conversation between one of the human players and Abel: "Abel, lo scopo del gioco è scoprire il codice nascosto formato da una sequenza di numeri. Le proposte sono 809, 752, 954, 830, 513. Ad ogni proposta corrisponde una risposta. Un

pallino nero rappresenta un numero giusto al posto giusto, quindi 809 al pallino nero, 752 al pallino nero, 513 al pallino nero. Mentre invece un pallino bianco indica la presenza di un numero giusto ma in posizione sbagliata, questo ce l'ha 830 e invece 954 non ha nulla.," transl. *Abel, the goal of the game is to discover the hidden code formed by a sequence of numbers. The guesses are 809, 752, 954, 830, 513. Each guess corresponds to a response. A black dot represents a correct number in the correct position, so 809 has a black dot, 752 has a black dot, 513 has a black dot. On the other hand, a white dot indicates the presence of a correct number but in the wrong position; 830 has a white dot, while 954 has nothing*. The conversation contains 5 sentences separated by a full stop. In addition, the distribution of features related to the syntactic tree structure of the texts suggests a more complex communication style when the conversation is with Abel rather than Box, as evidenced by the greater number of nominal complements, often organized into deeper embedded chains. Consider, for example, the following excerpt from a human-Abel conversation, "Qual è la capitale di Cipro del Nord?," transl. *What is the capital of northern Cyprus?*, where the noun "capitale," *capital*, is modified by a sequence of 2 embedded complement chains: "di Cipro," *of Cyprus*, modifies "capitale," *capital*, and "del Nord," lit. *of North*, modifies further "di Cipro," lit. *of Cyprus*.

Interestingly, the use of imperative verb forms is one of the highly statistically different features ($p < 0.01$) that characterizes higher levels of embodiment, as shown also in Table 9. In general, the group of features modeling verb inflectional morphology varies most when comparing conversations with Box and Avatar. Namely, on the one hand, when humans speak to Avatar, they tend to use a significantly higher number of second-person singular verbs (e.g., "Mi elenchi le isole del Mediterraneo orientale?," transl. *Can you list the islands in the eastern Mediterranean?*), while when talking with Box they use more first-person plural auxiliary verbs e.g., "Quale numero possiamo dire per risolvere questo enigma?," transl. *What number can we say to solve this riddle?*). Similarly to what was observed with Abel, higher levels of embodiment lead to more articulated utterances, characterized by a higher percentage of adjectives and adpositions (e.g., "in," *at*, "a," *to*), which typically introduce embedded chains of nominal complements, and of determiner relations holding between a nominal head and its determiner, including any words that modify a noun (e.g., "il," *the*, "questo," *this*, "quale," *which*), as well as case-marking relations used for any preposition introducing a noun, pronoun, adjective or adverb often occurring in a complement chain. In contrast, a feature that predominantly differentiates Box vs. Avatar human conversations is the use of punctuation marks. A detailed analysis of the punctuation distribution revealed that both Box and Avatar have the highest percentage of question marks, 0.34% and 0.36%, respectively, in relation to the overall punctuation count. On the other hand, as suggested by the statistical difference in punctuation marks between the Abel and Avatar conditions, our analysis showed that conversations with Abel involved the fewest number of question marks, indicating fewer questions. This dynamic may suggest that the Box and Avatar are more likely to be treated as tools for interrogation, while Abel, presenting the highest level of embodiment in our experimental setup, causes the conversation dynamics to shift toward a more dialogic style, reflecting possibly a higher sense of belonging [34].

After applying the FDR correction to control for multiple comparisons, as expected, only a subset of features remained statistically significant. These features can be interpreted as the most robust indicators of variation in the linguistic style of human participants when interacting with different artificial agents. Specifically, the number of sentences, reflecting the overall propensity of humans to interact with an artificial agent, and features modeling the lexical richness of utterances (*TTR:lemma-100 words, TTR:form-100 words*)[12] remain significant in the comparison between the

[12]Type–Token Ratio (TTR) refers to the ratio between the number of lexical types (considering both lemmas and word forms) and the number of tokens in a text. Due to its sensitivity to sample size, this feature is computed on text samples of equivalent length. In this case, it was calculated on the first 100 tokens of each message.

two most dissimilar agent embodiments (i.e., Box vs. Abel), while the use of proper nouns (*% proper nouns*), the trend toward greater referential specificity captured by the distribution of determiners (*% determiners*, both as morphosyntactic and syntactic feature), and the use of punctuation marks (*% punctuation*, as morphosyntactic feature), typically question marks, remain significant between conditions involving the most embodied agents (i.e., Abel vs. Avatar). Beyond these statistically robust effects, other linguistic features displayed patterns of variation that did not remain significant after correction. These patterns suggest meaningful yet less stable tendencies that may become more reliable with larger datasets, further supporting the interpretation that embodiment influences communicative behaviour in systematic ways.

5.2 Semantic Similarity across Conversations

In this analysis, we compute the SBERT score for each dialogue that takes place under different conditions involving an artificial player, complementing the previously introduced analysis of communication style. Since, as previously mentioned, the dialogue structure in the human-only condition differs significantly, we focus exclusively on human utterances, excluding messages from the artificial agents.

Specifically, we measured the SBERT score by comparing all transcribed messages in the pairs considered for the stylistic analysis, namely Box vs. Avatar, Box vs. Abel, and Avatar vs. Abel. For each session compared, we selected the maximum similarity score to obtain a single value, then averaged these maximum scores to provide an overall measure of semantic alignment across each escape room session for each experimental condition. As introduced in Section 3.2, the SBERT score ranges from -1 to 1. To quantify the content similarity of human dialogues across multiple conditions, we interpret a score of -1 as indicating completely dissimilar paired sentences, a score of 1 as indicating highly similar sentences, and a score of 0 as representing no meaningful similarity.

Figure 6 presents the comparison of the SBERT score across each artificial condition. The results indicate a moderate level of SBERT score similarity across all conditions and games, particularly between Box and the other two types of agent embodiments, Avatar and Abel. In both cases, the average cosine similarity is approximately 0.34, and we can reject the hypothesis—again using a *t-student* test for mean comparisons (see Footnote 7)—that it is equal to zero (p-value $< 0.1\%$). When comparing the dialogues between Abel and Avatar, we observe a higher degree of similarity with an average SBERT score of 0.4 (though still far from 1). These results are consistent with our initial hypothesis and with the findings reported for the linguistic style of conversation. Since the SBERT score quantifies the semantic similarity between pairs of texts by computing the cosine similarity between their sentence embeddings, we used this measure to estimate the degree of semantic alignment between utterances across experimental conditions. Accordingly, we expected that when the levels of embodiment are more similar (i.e., between Avatar and Abel), the content of human conversations would also be more similar, reflecting a greater degree of semantic coherence and alignment. This pattern aligns with the linguistic findings discussed above: a higher level of embodiment corresponds to fewer statistically varying linguistic features. Here, the higher stylistic similarity observed at the linguistic level is reflected in a higher semantic similarity in the conversational content. The results are robust across individual games, with the largest similarities appearing in Game 3 and the smallest in Game 4. This outcome confirms our intuition that the nature of Game 3, which relies on retrieving factual geographical knowledge to identify the city of Nicosia, constrains the dialogue to highly similar content across teams. Participants tend to ask comparable questions and seek the same kind of information to solve the game, independently of the agent's embodiment. In unreported non-parametric tests (e.g., Kolmogorov), we also check for significant differences in the distribution of cosine similarities across comparisons, but the results are consistent with the simple t-test analysis.

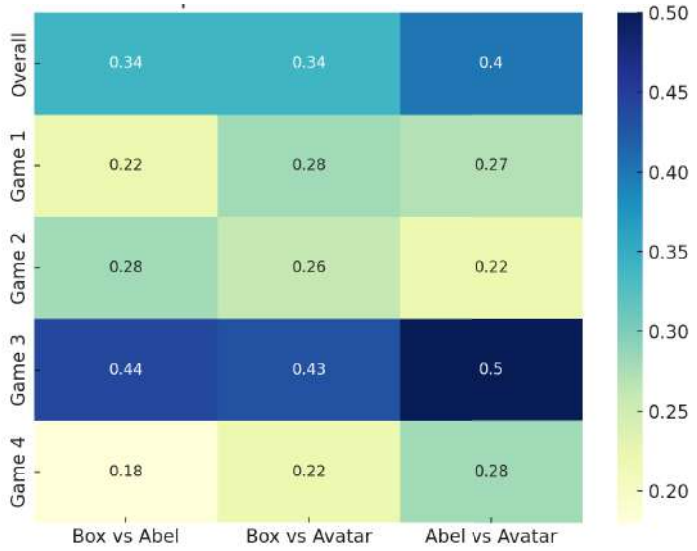


Fig. 6. SBERT score average similarity across paired experimental conditions.

Summary of Observed Communication Dynamics. As a general trend, we observed that teams composed only of humans interacted more frequently, producing a higher number of exchanged messages. This outcome is expected and aligns with the natural caution that speaking to an artificial agent may initially induce. However, our findings, though preliminary, suggest that the level of embodiment is related to the naturalness of human conversation. In fact, when interacting with the most embodied artificial agent, Abel, participants exchanged fewer messages overall, but their utterances were longer and exhibited features associated with a more natural communicative style. Moreover, we observed that human utterances addressed to artificial agents with more comparable levels of embodiment (i.e., Avatar and Abel) showed fewer differences in the number of statistically varying linguistic features, suggesting a greater convergence in communication style under higher embodiment conditions.

Higher embodiment thus appears to prompt human participants to adopt a communication style characterized by linguistic features suggesting that they treat the agent less as a tool to query for correct answers and more as a teammate to engage in interactive exchange. Specifically, when speaking to Abel, compared to the other artificial agents, participants tend to use a set of linguistic devices that all point to a more "natural" dialogic attitude. In particular, they more frequently use proper names to address Abel, morphologically mark verbs in the second-person singular and in the imperative form, and produce more syntactically articulated utterances (e.g., deeper nominal complement chains). Moreover, the transcriptions of their utterances contain fewer question marks. Importantly, the main patterns identified in this analysis remain reliable also after applying the FDR correction for multiple comparisons. Features that survive correction (i.e., the number of sentences, measures of lexical richness, and morpho-syntactic indicators including the distribution of determiners, proper nouns, and punctuation marks) represent the most robust evidence of linguistic variation across experimental settings.

The overall picture of the conversational dynamics is complemented by the analysis of the content of human conversations. Specifically, when the levels of embodiment of the artificial agents are more similar (i.e., between Avatar and Abel), the content of human conversations is also more similar, as reflected by higher SBERT scores. This similarity in conversational content aligns with

the stylistic findings reported above, reinforcing the interpretation that higher or more comparable levels of embodiment give rise not only to more natural linguistic forms but also to more aligned and coherent conversational content.

5.3 Discussion and Future Directions

Our findings contribute to the growing literature on non-dyadic human–robot and AI interaction by showing that embodiment systematically shapes how humans communicate with an artificial teammate and how this, in turn, relates to performance in non-routine analytical tasks. While prior HRI research has predominantly examined dyadic interactions [58], our results demonstrate that embodiment effects extend to collaborative group settings where coordination and shared reasoning are required. In particular, higher levels of embodiment elicit longer, richer, and more syntactically varied utterances toward the agent, indicating that humans adapt their communicative behaviour when the artificial teammate appears more socially and physically realistic. This pattern aligns with a growing body of evidence showing that anthropomorphic or socially expressive cues shape how humans communicate with artificial agents (e.g., [66]), especially within teams [36, 38, 47, 45, 63]. At the same time, we observe suggestive evidence of a non-linear pattern between the Avatar and Abel conditions, consistent with an uncanny-valley mechanism in which moderately realistic embodiment momentarily disrupts coordination before highly realistic embodiment restores fluency [25, 27, 41].

From a technical perspective, our findings resonate with recent work documenting the challenges robots face in detecting group structure, managing turn-taking, and monitoring engagement in multiparty settings [51]. Higher embodiment in our study appeared to partially compensate for these perceptual limitations: the agent's physical realism elicited clearer and more sustained human–agent communication, even though the AI itself was not actually perceiving or understanding the conversation any better. In this sense, the presence of a more embodied agent shaped the team's overall communicative climate, reinforcing the view that group HRI functions as a socio-technical system in which social cues guide cognitive effort [60]. Our findings also point to the need to test different underlying AI architectures. The agent in our study relied on a specific language model, but alternative models, with richer grounding or more robust dialogue and multimodal abilities, may produce different communication patterns [4].

From an economic perspective, our work also aligns with prior research on generative AI and collaborative work in innovative tasks [22, 54], which shows that artificial agents often have heterogeneous effects: they help some users or teams while hindering others, increasing performance variance rather than producing uniform gains [54]. Our results reflect this pattern. Embodied agents did not uniformly improve outcomes; instead, they increased dispersion in performance. Some teams coordinated more rapidly with embodied agents, while others struggled to integrate the artificial partner, generating an uncanny-valley-like increase in variance: dispersion widens even further at moderate embodiment levels (the Avatar) and contracts only when realism is high (Abel), yet overall remains larger than in human-only teams. This highlights the need for future HRI research to model not only mean effects but also the distributional consequences of embodiment.

A further implication that deserves more attention concerns the role of errors in non-routine problem solving. Human-only teams made more errors, yet were more likely to ultimately escape the room. This pattern is consistent with research on creativity and innovation showing that exploratory mistakes are integral to complex problem solving, reflecting hypothesis testing, cognitive flexibility, and divergent search [39]. Creative teams often progress through cycles of trial, error, and refinement rather than linear optimisation. Our findings align with this view: making more errors was associated with broader exploration and, ultimately, higher success. Mixed teams, guided by the AI agent,

tended to make fewer errors, suggesting a more conservative or AI-led strategy that may limit exploratory search.

We also acknowledge that Abel's youthful, gender-neutral appearance may have influenced participant expectations or social attributions. Research shows that social identity cues and perceived vulnerability shape interaction norms, sometimes eliciting more instructive, guiding, or protective behaviour toward robots that appear socially dependent or vulnerable [36, 63]. While we did not directly manipulate perceived age, this remains an important dimension for future work examining how embodiment interacts with role expectations and status cues in mixed human–AI teams [34]. Future research should also assess these dynamics in more diverse populations, as sensitivity to perceived age, gender cues, and social roles may vary across demographic groups and cultural backgrounds (e.g., [50]). A related limitation concerns the measurement of embodiment itself. Although the three conditions were defined *ex ante* by the agents' physical and visual features, we did not administer a validated embodiment or anthropomorphism scale to assess participants' perceptions directly. We collected a single post-experimental human-likeness item, but it does not significantly differ across conditions and may be too limited for modern language-model-based agents, whose perceived humanness may also depend on deeper dimensions such as responsiveness, fluency, competence, and social presence [12]. Future studies should therefore develop and adopt richer measures of embodiment and anthropomorphism [59].

Methodologically, our study develops an innovative experimental setup inspired by [32] work on human-only teams, extending it to mixed human–AI collaboration. The design is fully reproducible: all puzzles, instructions, solution logic, hint structures, and timing rules are documented in Appendix, enabling straightforward replication. The framework is also modular and easily adjustable. Researchers can tailor difficulty by adding or removing hints or by substituting entire tasks, for example, replacing linguistic, numerical, or code-breaking challenges with different ones, while preserving the core structure of the experiment. At the same time, the evidence based on our questionnaire controls remains exploratory and correlational. In particular, the suggestive pattern linking higher entropy in game preferences to better performance should be interpreted cautiously, as it may reflect within-team heterogeneity in skills or preparedness rather than a validated mechanism. Future work should examine this dimension more directly using richer measures and, where possible, exogenous variation in team competences as in [67]. As with any costly field or field-in-the-lab design involving highly sophisticated artificial agents, our study faces natural limits in statistical power. This reinforces the value of coordinated multi-lab replication efforts, similar to replication initiatives in experimental economics, to scale up evidence on human–AI teamwork and to strengthen the external validity of HRI research.

Along this line, future research can be extended in several directions. First, the potential non-linear relationship between embodiment and coordination—where moderate embodiment (Avatar) may introduce ambiguity, while highly realistic embodiment (Abel) facilitates fluency—deserves deeper theoretical and computational investigation. Second, longitudinal studies could examine whether communication adaptation stabilises as teams become more familiar with the agent or whether novelty effects fade. Third, cross-cultural studies could identify how norms of politeness, hierarchy, or uncertainty avoidance shape embodiment effects.

6 Conclusions

In this article, we examined how artificial agents with different levels of embodiment shape team performance and communication in a real-life escape-room task requiring non-routine analytical reasoning. Teams collaborated either exclusively with other humans or with an artificial teammate embodied as a Box, an Avatar, or the hyper-realistic humanoid Abel. Across four heterogeneous challenges, we evaluated success rates, completion times, errors, and linguistic dynamics.

Our findings reveal a counterintuitive performance pattern. On average, human-only teams were more likely to complete all tasks within the allotted time, while teams with artificial agents, especially more embodied ones, tended to commit fewer errors and were, in some cases, either much faster or considerably slower. Embodiment also influenced communication: interactions with our hyper-realistic humanoid Abel elicited longer, more natural, and more varied utterances, indicating that humans align their communication style to the agent's physical and social realism. These results suggest that the level of embodiment in artificial agents not only influences team performance but also shapes how humans engage in dialogue during collaborative problem-solving tasks and increases performance variance across teams. This heterogeneous pattern is consistent with recent evidence on the uneven effects of generative AI in collaborative tasks [54] and also aligns with a general uncanny-valley dynamic, in which intermediate levels of realism may momentarily hinder coordination before higher realism restores fluency.

Finally, we find that heterogeneity within teams matters. Higher entropy in team members' game preferences and competencies was associated with better performance, suggesting that diverse problem-solving approaches and complementary strengths play a meaningful role in complex group tasks. This highlights the need for future work to consider not only average treatment effects, but also the distributional consequences of embodiment and the role of within-team diversity in shaping human–AI collaboration.

Although still exploratory, this study provides convergent evidence that artificial teammates reshape the early stages of interaction within human teams, influencing both communication patterns and performance. By releasing our full dataset and detailed experimental materials, we aim to support cumulative research on multiparty human–AI collaboration and enable replications and extensions.

Building on this work, we are indeed moving to a new data collection with an enhanced perceptual architecture. The new setup incorporates improved real-time speech detection, multimodal object recognition, and collecting physiological monitoring (heart rate, electrodermal activity), enabling us to examine how emotional arousal and stress evolve during collaboration with embodied AI. These developments will allow us to study affective and cognitive alignment in human–AI teams and to design adaptive, emotionally responsive artificial agents capable of sustaining effective interaction in complex group environments.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process. During the preparation of this work, the authors used ChatGPT-4 in order to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Acknowledgments

The authors dedicate this work to Prof. Danilo Emilio De Rossi, without whom ABEL, and indeed our entire research team, would not have come into being. His vision and guidance have been, and will remain, a constant source of inspiration.

The authors also thank Simeon Schudy, participants at the Machine Behavior Conference 2024 in Berlin, for insightful comments, and Lorenzo Cascone and Davide Borghini for excellent research assistance.

References

[1] Daron Acemoglu, David Autor, and Simon Johnson. 2023. Can We Have Pro-Worker AI? Choosing a path of machines in service of minds. In *CEPR Policy Insight 123*. CEPR Press. Retrieved from https://cepr.org/publications/policy-insight-123-can-we-have-pro-worker-ai-choosing-path-machines-service-minds

[2] Achyuta Adhvaryu, Namrata Kala, and Anant Nyshadham. 2023. Returns to on-the-job soft skills training. *Journal of Political Economy* 131, 8 (2023), 2165–2208.

[3] Safinah Ali, Nisha Elizabeth Devasia, and Cynthia Breazeal. 2022. Escape! bot: Social robots as creative problem-solving partners. In *Proceedings of the 14th Conference on Creativity and Cognition*, 275–283.
[4] Jesse Atuhurra. 2024. Large language models for human-robot interaction: Opportunities and risks. arXiv:2405.00693. Retrieved from https://arxiv.org/abs/2405.00693
[5] David Autor. 2024. *Applying AI to Rebuild Middle Class Jobs*. Technical Report. National Bureau of Economic Research.
[6] Sarah Bankins, Anna Carmella Ocampo, Mauricio Marrone, Simon Lloyd D. Restubog, and Sang Eun Woo. 2024. A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *Journal of Organizational Behavior* 45, 2 (2024), 159–182.
[7] Daniel Baumgartner and John Kolassa. 2023. Power considerations for kolmogorov–Smirnov and Anderson–Darling two-sample tests. *Communications in Statistics - Simulation and Computation* 52, 7 (2023), 3137–3145.
[8] Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F. Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L. Griffiths, Joseph Henrich, et al. 2023. Machine culture. *Nature Human Behaviour* 7, 11 (2023), 1855–1868.
[9] Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: A tool for linguistic profiling of texts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 7145–7151.
[10] Jason W. Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A. Bakker, Joshua A. Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, et al. 2024. How large language models can reshape collective intelligence. *Nature Human Behaviour* 8 (2024), 1643–1655.
[11] Andrew Caplin, David J. Deming, Shangwen Li, Daniel J. Martin, Philip Marx, Ben Weidmann, and Kadachi Jiada Ye. 2024. *The ABC's of Who Benefits from Working with AI: Ability, Beliefs, and Calibration*. Technical Report. National Bureau of Economic Research.
[12] Oscar Hengxuan Chi, Christina G. Chi, and Dogan Gursoy. 2025. Seeing personhood in machines: Conceptualizing anthropomorphism of social robots. *Journal of Service Research* 28, 1 (2025), 78–92.
[13] M. Chugunova and D. Sele. 2022. We and it: An interdisciplinary review of the experimental evidence on how humans interact with machines. *Journal of Behavioral and Experimental Economics* 99 (2022), 101897.
[14] Lorenzo Cominelli, Francesco Feri, Roberto Garofalo, Caterina Giannetti, Miguel A. Meléndez-Jiménez, Alberto Greco, Mimma Nardelli, Enzo Pasquale Scilingo, and Oliver Kirchkamp. 2021. Promises and trust in human–robot interaction. *Scientific Reports* 11, 1 (2021), 1–14.
[15] Lorenzo Cominelli, Gustav Hoegen, and Danilo De Rossi. 2021. Abel: Integrating humanoid body, emotions, and time perception to investigate social interaction and human cognition. *Applied Sciences* 11, 3 (2021), 1070.
[16] Brice Corgnet, Roberto Hernán-González, and Ricardo Mateo. 2023. Peer effects in an automated world. *Labour Economics* 85 (2023), 102455.
[17] Filipa Correia, Samuel Mascarenhas, Rui Prada, Francisco S. Melo, and Ana Paiva. 2018. Group-based emotions in teams of humans and robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 261–269.
[18] Vittorio Cuculo and Alessandro D'Amelio. 2019. OpenFACS: An open source FACS-based 3D face animation system. In *Image and Graphics*. Yao Zhao, Nick Barnes, Baoquan Chen, Rüdiger Westermann, Xiangwei Kong, and Chunyu Lin (Eds.), Springer International Publishing, Cham, 232–242.
[19] Hao Cui and Taha Yasseri. 2024. AI-enhanced collective intelligence. *Patterns* 5 (2024), 11.
[20] Julian De Freitas, Stuti Agarwal, Bernd Schmitt, and Nick Haslam. 2023. Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour* 7, 11 (2023), 1845–1854.
[21] Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47, 2 (2021), 255–308.
[22] Fabrizio Dell'Acqua, Charles Ayoubi, Hila Lifshitz, Raffaella Sadun, Ethan Mollick, Lilach Mollick, Yi Han, Jeff Goldman, Hari Nair, Stewart Taub, et al. 2025. *The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork And Expertise*. Technical Report. National Bureau of Economic Research.
[23] Fabrizio Dell'Acqua, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality*. Harvard Business School Technology & Operations Mgt. Unit Working Paper 24-013. Harvard Business School.
[24] Mustafa Demir, Nathan J. McNeese, and Nancy J. Cooke. 2020. Understanding human-robot teams in light of all-human teams: Aspects of team interaction and shared cognition. *International Journal of Human-Computer Studies* 140 (2020), 102436.
[25] Matthieu Destephe, Martim Brandao, Tatsuhiro Kishi, Massimiliano Zecca, Kenji Hashimoto, and Atsuo Takanishi. 2015. Walking in the uncanny valley: Importance of the attractiveness on the acceptance of a robot as a working partner. *Frontiers in Psychology* 6 (2015), 204.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805
[27] Alexander Diel, Sarah Weigelt, and Karl F. Macdorman. 2021. A meta-analysis of the uncanny valley's independent and dependent variables. *ACM Transactions on Human-Robot Interaction* 11, 1 (2021), 1–33.
[28] Virginia Dignum. 2022. Relational artificial intelligence. arXiv:2202.07446. Retrieved from https://arxiv.org/abs/2202.07446
[29] Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association* 9, 3 (2011), 522–550.
[30] Wen Duan, Shiwen Zhou, Matthew J. Scalia, Xiaoyun Yin, Nan Weng, Ruihao Zhang, Guo Freeman, Nathan McNeese, Jamie Gorman, and Michael Tolston. 2024. Understanding the evolvement of trust over time within human-AI teams. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–31.
[31] Christoph Engel, Max, R. P. Grossmann, and Axel Ockenfels. 2024. *Integrating Machine Behavior into Human Subject Experiments: A User-Friendly Toolkit and Illustrations*. MPI Collective Goods Discussion Paper 2024/1. Max Planck Institute for Research on Collective Goods. Retrieved from https://pure.mpg.de/rest/items/item_3558922_3/component/file_3558923/content
[32] Florian Englmaier, Stefan Grimm, Dirk Schindler, and Simeon Schudy. 2024. The effect of incentives in non-routine analytical teams tasks-evidence from a field experiment. *Journal of Political Economy* 132, 8 (2024), 2695–2747.
[33] Sonja Engmann and Denis Cousineau. 2011. Comparing distributions: The two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test. *Journal of Applied Quantitative Methods* 6, 3 (2011), 1–17.
[34] Hadas Erel, Marynel Vázquez, Sarah Sebo, Nicole Salomons, Sarah Gillet, and Brian Scassellati. 2024. RoSI: A model for predicting robot social influence. *ACM Transactions on Human-Robot Interaction* 13, 2 (2024), 1–22.
[35] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467.
[36] Marlena R. Fraune. 2020. Our robots, our team: Robot anthropomorphism moderates group effects in human–robot teams. *Frontiers in Psychology* 11 (2020), 1275.
[37] Matthew C. Gombolay, Reymundo A. Gutierrez, Shanelle G. Clarke, Giancarlo F. Sturla, and Julie A. Shah. 2015. Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots* 39 (2015), 293–312.
[38] Alexandra M. Harris-Watson, Lindsay E. Larson, Nina Lauharatanahirun, Leslie A. DeChurch, and Noshir S. Contractor. 2023. Social perception in Human-AI teams: Warmth and competence predict receptivity to AI teammates. *Computers in Human Behavior* 145 (2023), 107765.
[39] Sarah Harvey. 2014. Creative synthesis: Exploring the process of extraordinary group creativity. *Academy of Management Review* 39, 3 (2014), 324–343.
[40] Natalia Kalashnikova, Mathilde Hutin, Ioana Vasilescu, and Laurence Devillers. 2023. Do we speak to robots looking like humans as we speak to humans? A study of pitch in French human-machine and human-human interactions. In *Companion Publication of the 25th International Conference on Multimodal Interaction (ICMI '23 Companion)*, 141–145.
[41] Hangyeol Kang, Thiago Freitas dos Santos, Maher Ben Moussa, and Nadia Magnenat-Thalmann. 2025. Mitigating the uncanny valley effect in hyper-realistic robots: A student-centered study on LLM-driven conversations. arXiv:2503.16449. Retrieved from https://arxiv.org/abs/2503.16449
[42] Theodora Koulouri, Stanislao Lauria, and Robert D. Macredie. 2016. Do (and say) as I say: Linguistic adaptation in human–computer dialogs. *Human–Computer Interaction* 31, 1 (2016), 59–95.
[43] Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies* 77 (2015), 23–37.
[44] Ning Li, Huaikang Zhou, and Kris Mikel-Hong. 2024. Generative AI enhances team performance and reduces need for traditional teams. arXiv:2405.17924. Retrieved from https://arxiv.org/abs/2405.17924
[45] Rohit Mallick, Christopher Flathmann, Wen Duan, Beau G. Schelble, and Nathan J. McNeese. 2024. What you say vs. what you do: Utilizing positive emotional expressions to relay AI teammate intent within human–AI teams. *International Journal of Human-Computer Studies* 192 (2024), 103355.
[46] Thomas W. Malone and Michael Bernstein. 2022. *Handbook of Collective Intelligence*. MIT Press.
[47] 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.
[48] Nathan J. McNeese, Christopher Flathmann, Thomas A. O'Neill, and Eduardo Salas. 2023. Stepping out of the shadow of human-human teaming: Crafting a unique identity for human-autonomy teams. *Computers in Human Behavior* 148 (2023), 107874.

[49] Nathan J. McNeese, Beau G. Schelble, Lorenzo Barberis Canonico, and Mustafa Demir. 2021. Who/what is my teammate? Team composition considerations in human–AI teaming. *IEEE Transactions on Human-Machine Systems* 51, 4 (2021), 288–299.

[50] Lucas Morillo-Mendez, Martien G. S. Schrooten, Amy Loutfi, and Oscar Martinez Mozos. 2024. Age-related differences in the perception of robotic referential gaze in human-robot interaction. *International Journal of Social Robotics* 16, 6 (2024), 1069–1081.

[51] Massimiliano Nigro, Emmanuel Akinrintoyo, Nicole Salomons, and Micol Spitale. 2025. Social group human-robot interaction: A scoping review of computational challenges. In *2025 Proceedings of the 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 468–478.

[52] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (2023), 187–192.

[53] Thomas A. O'Neill, Christopher Flathmann, Nathan J. McNeese, and Eduardo Salas. 2023. 21st century teaming and beyond: Advances in human-autonomy teamwork. *Computers in Human Behavior* 147 (2023), 107865.

[54] Nicholas Otis, Rowan, P. Clarke, Solene Delecourt, David Holtz, and Rembrand Koning. 2023. The Uneven Impact of Generative AI on Entrepreneurial Performance. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4671369

[55] Dino Pedreschi, Luca Pappalardo, Ricardo Baeza-Yates, Albert-Laszlo Barabasi, Frank Dignum, Virginia Dignum, Tina Eliassi-Rad, Fosca Giannotti, Janos Kertesz, Alistair Knott, et al. 2023. Social AI and the challenges of the human-AI ecosystem. arXiv:2306.13723. Retrieved from https://arxiv.org/abs/2306.13723

[56] Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The impact of AI on developer productivity: Evidence from GitHub Copilot. arXiv:2302.06590. Retrieved from https://arxiv.org/abs/2302.06590

[57] N, Reimers. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv:1908.10084. Retrieved from https://arxiv.org/abs/1908.10084

[58] Eike Schneiders, EunJeong Cheon, Jesper Kjeldskov, Matthias Rehm, and Mikael B. Skov. 2022. Non-dyadic interaction: A literature review of 15 years of human-robot interaction conference publications. *ACM Transactions on Human-Robot Interaction* 11, 2 (2022), 1–32.

[59] Mariah Schrum, Muyleng Ghuy, Erin Hedlund-Botti, Manisha Natarajan, Michael Johnson, and Matthew Gombolay. 2023. Concerning trends in likert scale usage in human-robot interaction: Towards improving best practices. *ACM Transactions on Human-Robot Interaction* 12, 3 (2023), 1–32.

[60] Sarah Sebo, Brett Stoll, Brian Scassellati, and Malte F. Jung. 2020. Robots in groups and teams: A literature review. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–36.

[61] Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC '16)*. Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, and Jan Odijk and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Portorož, Slovenia, 4290–4297. Retrieved from https://aclanthology.org/L16-1680

[62] Hamish Tennent, Solace Shen, and Malte Jung. 2019. Micbot: A peripheral robotic object to shape conversational dynamics and team performance. In *Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 133–142.

[63] Margaret L. Traeger, Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A. Christakis. 2020. Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences* 117, 12 (2020), 6370–6375.

[64] Milena Tsvetkova, Taha Yasseri, Niccolo Pescetelli, and Tobias Werner. 2024. A new sociology of humans and machines. *Nature Human Behaviour* 8, 10 (2024), 1864–1876.

[65] Hans van Halteren. 2004. Linguistic profiling for author recognition and verification. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 200–207.

[66] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology* 52 (2014), 113–117.

[67] Ben Weidmann and David J. Deming. 2021. Team players: How social skills improve team performance. *Econometrica* 89, 6 (2021), 2637–2657.

[68] Hans W. Wendt. 1972. Dealing with a common problem in social science: A simplified rank-biserial coefficient of correlation based on the statistic. *European Journal of Social Psychology* 2, 4 (1972), 463–465.

[69] F. Wolf and R. Stock-Homburg. 2023. How and when can robots be team members? Three decades of research on human–robot teams. *Group & Organization Management* 48, 6 (2023), 1666–1744.

[70] Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, and Iyad Rahwan. 2024. Empirical evidence of large language model's influence on human spoken communication. arXiv:2409.01754. Retrieved from https://arxiv.org/abs/2409.01754

[71] Rui Zhang, Wen Duan, Christopher Flathmann, Nathan McNeese, Guo Freeman, and Alyssa Williams. 2023. Investigating AI teammate communication strategies and their impact in human-AI teams for effective teamwork. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–31.

[72] Rui Zhang, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2021. "An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.

Appendix

Escape Room Description

In total, four games must be solved sequentially to exit ("escape") the room. The team's final objective is to find the key hidden inside a book on the library shelves to the left, which is locked with a secret combination that must be discovered as the last of the four games. Hidden hints (pieces of paper scattered around the room) can be found in two ways: either by searching the room—visually or by moving objects—or by solving a game, which provides useful information to locate a slip of paper (see Figure A1). Each hint helps solve the next game. For example, Hint #1 will help solve Game 2. Therefore, the only game without a hint is the first one.

Game 1: Identifying the Odd Word. The first game requires selecting the one word out of seven that does not belong to a list. Six words are anagrams of sports, while the seventh is an anagram of the Italian word "table," where the first hint is hidden. This hint reads: "The solution is literally the number."

Game 2: The Secret Password. In the second game, players must enter a secret password to access "a bandits" hideout." Mathematically, one might be tempted to assume that the password follows the rule "half the number," but instead, the correct rule is "count the number of letters." Since the correct answer has seven letters, the second hint is hidden near a chessboard on the right table, with seven pieces placed outside. It reads: "The last divided capital."

Table A1. Summary Statistics

Variable	Mean	SD	Min	Max	Description
Age 20	0.10	0.30	0.00	1.00	Dummy variable equal to 1 if the participant is aged 20 or younger, 0 otherwise.
Age 21–25	0.50	0.50	0.00	1.00	Dummy variable equal to 1 if the participant is aged between 21 and 25 years, 0 otherwise.
Age 26–30	0.29	0.46	0.00	1.00	Dummy variable equal to 1 if the participant is aged between 26 and 30 years, 0 otherwise.
Age 31+	0.11	0.31	0.00	1.00	Dummy variable equal to 1 if the participant is older than 30 years, 0 otherwise.
STEM field	0.44	0.50	0.00	1.00	Dummy variable equal to 1 if the field of study is in STEM disciplines, 0 otherwise.
Regularly use AI	4.56	2.09	1.00	7.00	Frequency of AI tool use on a 1–7 Likert item (higher = more frequent use).
Tech curiosity	3.41	1.74	1.00	7.00	Self-reported curiosity about new technologies (1–7 Likert item).
Tech propensity	5.44	1.47	1.00	7.00	Propensity to use and adopt technology (1–7 Likert item).
Risk propensity	2.22	0.80	1.00	4.00	Willingness to take risks, measured on a 1–4 Likert item.
Leader	2.75	0.85	1.00	4.00	Self-assessed leadership level (1–4 Likert item, higher = stronger leadership).
Control preference	2.44	0.88	1.00	4.00	Preference for having control over group decisions (1–4 Likert item).
Favorite game 1	0.21	0.41	0.00	1.00	Dummy variable equal to 1 if the participant's favorite game is Type 1, 0 otherwise.
Favorite game 2	0.18	0.38	0.00	1.00	Dummy variable equal to 1 if the participant's favorite game is Type 2, 0 otherwise.
Favorite game 3	0.26	0.44	0.00	1.00	Dummy variable equal to 1 if the participant's favorite game is Type 3, 0 otherwise.
Favorite game 4	0.25	0.44	0.00	1.00	Dummy variable equal to 1 if the participant's favorite game is Type 4, 0 otherwise.
No favorite game	0.10	0.30	0.00	1.00	Dummy variable equal to 1 if the participant reported no favorite game, 0 otherwise.
Entropy game	0.67	0.18	0.00	1.00	Normalized Shannon of favorite game choices within each group (1 = maximally diverse).
Escape experience	0.36	0.48	0.00	1.00	Dummy variable equal to 1 if the participant has prior escape room experience, 0 otherwise.
Group friendship	3.13	2.20	1.00	7.00	Perceived level of friendship among group members (1–7 Likert item).
Group connection	5.61	1.46	1.00	7.00	Degree of connection and rapport within the group (1–7 Likert item).
Group Conflicts	1.37	1.06	1.00	7.00	Frequency of conflicts experienced within the group (1–7 Likert item).
Group harmony	6.02	1.21	1.00	7.00	Perceived harmony and cooperation within the group (1–7 Likert item).
Observations	174				

Group interaction variables include friendships, connection, conflicts, and harmony (1–7). Individual characteristics cover demographics, field of study, technology attitude (1–7), risk propensity (1–4), leadership (1–4), control preference (1–4), AI perception (1–7), and escape room experience (Yes/No). The entropy variable quantifies the diversity of favorite game types within each session using the normalized Shannon index.

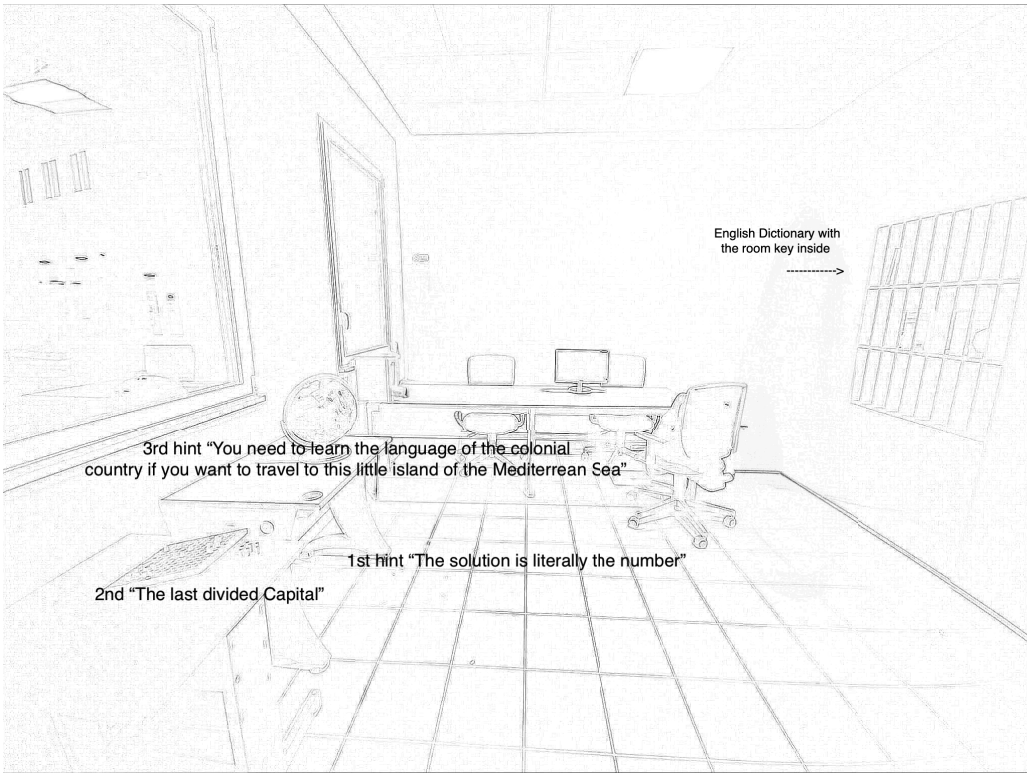


Fig. A1. Escape room overview.

Game 3: The Divided Capital. Participants must identify the city of Nicosia from a picture of the Turkish-Greek border. They must also guess the country where the city is located. Solving this puzzle leads to the third hint, hidden under a globe in the room. It reads "This language you need to learn if you want to go to this small island of the colonial empire."

Game 4: The Final Puzzle. The last clue should direct players to "The New English Dictionary" on the shelf. Here, participants must solve a master puzzle to open the book and enter the solution on the computer—which ultimately unlocks the key and allows them to escape the room.

Instructions for Participants

Treatment Only-Humans. Addressed to the 3 participants: "*In this game, you will form a team and work together to complete it.*"

Treatment ABEL-Humans. Addressed to the 2 human participants: "*In this game, you will form a team together with the humanoid ABEL, which you see in front of you, and work together to complete it.*"

Treatment Avatar-Humans. Addressed to the 2 human participants: "*In this game, you will form a team together with the avatar ABEL, which is now visible on the screen, and work together to complete it.*"

Treatment Box-Humans. Addressed to the 2 human participants: "*In this game, you will form a team together with the artificial agent, which you see in front of you, and work together to complete it.*"

The goal of the game is to escape the room in the shortest possible time, and in any case within 25 minutes. The game ends when you find the key to exit the room and enter, on the computer in

front of you on the table, the answer indicating the physical location of the key. Regardless of the outcome of the game—whether you manage to exit the room or not—you will receive a voucher to collect a T-shirt from the University of Pisa at the university store. Additionally, in a few weeks, you will receive an e-mail with the partial rankings. If your group is the fastest every 25 groups, each member will receive a hoodie from the University of Pisa. To find the key and escape the room, you must solve 4 games, which are interconnected. You cannot solve the fourth game, and therefore, find the key, without first completing the third. Similarly, the third game depends on the second, and the second on the first. To help you solve the games, there are 3 hints hidden in the room, represented by numbered slips of paper. You can find these hints in two ways:

(1) Searching the room, visually or by moving objects.
(2) Solving a game, which will provide useful information to locate a slip of paper.

Each hint will help you solve the next game. For example, hint #1 will help you solve game 2. Therefore, the only game without hints is the first one. It is important that you use the personal microphones, which we will now have you wear, to communicate with each other. This is followed by an explanation of the equipment, including a sample screen showing how to read the question, enter answers, and verify their correctness to proceed to the next phase. At this stage, communication with the artificial agent will also be tested.

Survey Questions (Translated to English)

(1) Age: How old are you? (15–99)
(2) Gender: What is your gender?
(3) Field of Study: What is your field of study?
(4) I generally prefer to be a leader rather than a follower. (Response format: 1–4)
(5) I try to avoid situations where others tell me what to do. (Response format: 1–4)
(6) Generally, in life, I tend to avoid taking risks. (Response format 1–4)
(7) How much would you say the artificial agent is human-like? (Response format 1–7)
(8) I enjoy using technological systems in depth. (Response format 1–7)
(9) It is enough for me that a technological system works; I do not care how or why. (Response format 1–7)
(10) I am used to interacting with generative AI systems (e.g., ChatGPT). (Response format 1–7)
(11) How much would you say the other group members are your friends? (Response format 1–7)
(12) How much would you say there was team synergy during the game? (Response format 1–7)
(13) How much would you say there were conflicts during the game? (Response format 1–7)
(14) Have you ever played an escape room before today? (Yes/No)
(15) Which game made you feel most prepared to play? (Options: 1, 2, 3, 4, None)
(16) How much would you say there was harmony and collaboration in the group during the game? (Response format: 1–7)

Received 7 November 2024; revised 20 March 2026; accepted 16 April 2026