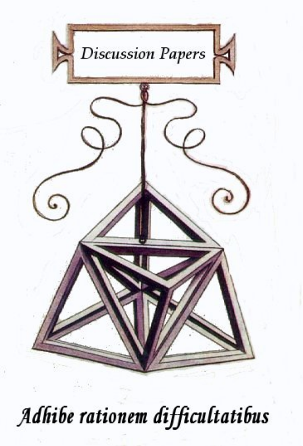
Discussion papers
E-papers of the Department of Economics e Management – University di Pisa



Lorenzo Cominelli, Federico Galatolo,
Caterina Giannetti, Cristiano Ciaccio,
Felice Dell'Orletta, Philipp Chaposkvi,
Giulia Venturi

Teaming Up with Artificial Agents in Non-routine Analytical Tasks

Discussion paper n. 314

2024

Discussion paper n. 314, presented: **November 2024**

Authors' address/Indirizzo degli autori:

Lorenzo Cominelli — University of Pisa - Information Engineering Dept.- Largo Lucio Lazzarino 1. E-mail: lorenzo.cominelli@unipi.it

Federico Galatolo — University of Pisa -Information Engineering Dept.- Largo Lucio Lazzarino 1. E-mail: federico.galatolo@unipi.it

Caterina Giannetti — University of Pisa, Department of Economics and Management. E-mail: caterina.giannetti@gmail.com

Cristiano Ciaccio — Institute of Computational Linguistics "Antonio Zampolli", Pisa, Italy. E-mail: cristiano.ciaccio@ilc.cnr.it

Felice Dell'Orletta — Institute of Computational Linguistics "Antonio Zampolli", Pisa, Italy. E-mail: felice.dellorletta@ilc.cnr.it

Philipp Chaposkvi — Universitat Duisburg Essen Institut fur Politikwissenschaft. E-mail: chapkovski@gmail.com

Giulia Venturi — Institute of Computational Linguistics "Antonio Zampolli", Pisa, Italy. E-mail: giulia.venturi@ilc.cnr.it

Please cite as:/Si prega di citare come:
Lorenzo Cominelli, Federico Galatolo, Caterina Giannetti, Cristiano Ciaccio, Felice Dell'Orletta, Philipp Chaposkvi, Giulia Venturi (2024), "Teaming Up with Artificial Agents in Non-routine Analytical Tasks", Discussion Papers, Department of Economics and Management – University of Pisa, n. 314 (http://www.ec.unipi.it/ricerca/discussion-papers).

Discussion Papers Series contact: pietro.battiston@unipi.it

Discussion paper

n. 314

Discussion Papers

Adhibe rationem difficultatibus

Lorenzo Cominelli, Federico Galatolo, Caterina Giannetti, Cristiano Ciaccio, Felice Dell'Orletta, Philipp Chaposkvi, Giulia Venturi

Teaming Up with Artificial Agents in Non-routine Analytical Tasks

Abstract

Using a real-life escape room scenario, we investigate how different levels of embodiment in artificial agents influence team performance and conversational dynamics in non-routine analytical tasks. Teams composed of either three humans or two humans and an artificial agent (a Box, an Avatar, and a hyper-realistic Humanoid) worked together to escape the room within a time limit. Our findings reveal that while human-only teams tend to complete all tasks more frequently, they also tend to be slower and make more errors. Additionally, we observe a non-linear relationship between the degree of agent embodiment and team performance, with a significant effect on conversational dynamics. Teams with agents exhibiting higher levels of embodiment display conversational patterns more similar to those occurring among humans. These results highlight the complex role that embodied AI plays in human-agent interactions, offering new insights into how artificial agents can be designed to support team collaboration in problem-solving environments.

Keywords: complex tasks, artificial agents, teamwork

JEL CLassification: C92

Teaming Up with Artificial Agents in Non-routine Analytical Tasks

Lorenzo COMINELLI[1], Federico GALATOLO[1,2], Caterina GIANNETTI[2,3] *
Cristiano CIACCIO[4], Felice DELL'ORLETTA[4], Philipp CHAPKOVSKI[5]
Giulia VENTURI[4]

October 31, 2024

Abstract

Using a real-life escape room scenario, we investigate how different levels of embodiment in artificial agents influence team performance and conversational dynamics in non-routine analytical tasks. Teams composed of either three humans or two humans and an artificial agent (a Box, an Avatar, and a hyper-realistic Humanoid) worked together to escape the room within a time limit. Our findings reveal that while human-only teams tend to complete all tasks more frequently, they also tend to be slower and make more errors. Additionally, we observe a non-linear relationship between the degree of agent embodiment and team performance, with a significant effect on conversational dynamics. Teams with agents exhibiting higher levels of embodiment display conversational patterns more similar to those occurring among humans. These results highlight the complex role that embodied AI plays in human-agent interactions, offering new insights into how artificial agents can be designed to support team collaboration in problem-solving environments.

*[1]Department of Information Engineering, University of Pisa
[2]Department of Economics and Management, University of Pisa, Italy, Corresponding author: Email caterina.giannetti@unipi.it
[3] Center E. Piaggio, University of Pisa, Italy
[4]Institute of Computational Linguistics "Antonio Zampolli", Pisa, Italy.
[5] Universitat Duisburg Essen Institut fur Politikwissenschaft

The authors acknowledge support from the project "Teaming-up with social artificial agents" funded by the Program for Research Projects of National Interest (PRIN) grant no 2022ALB-SWX. The authors also thank Simeon Schudy, participants at the Machine Behavior Conference 2024 in Berlin, for insightful comments, and Lorenzo Cascone for excellent research assistance.

1 Introduction

As workplaces increasingly integrate automation in daily routines, understanding the dynamics of human collaboration with artificial intelligent (AI) agents has become crucial. However, research on the integration of AI technologies into human teams remains scarce and yields mixed results (Johnson et al., 2012; Sebo et al., 2020; Li et al., 2024). Indeed, inconsistencies are largely driven because the type of task plays a critical role in shaping outcomes. For example, when AI agents perform tasks that are similar to their human counterparts, such as style production tasks, their presence can diminish human performance due to reduced social incentives and peer effects (Corgnet et al., 2023). In contrast, when tasks are distinct, with AI agents taking on different roles than humans, there is often an increase in human effort, possibly because the AI agents are seen as complementary rather than competitive (Gombolay et al., 2015; Li et al., 2024). The complexity of tasks, particularly when they are open-ended and require creativity or interpersonal collaboration, is also crucial in shaping outcomes. Evidence shows that AI support in complex tasks can increase variance in performance, benefiting high performers more (Otis et al., 2023). In structured environments, however, AI tends to enhance productivity and quality, particularly for lower performers, though task complexity still plays a role in determining effectiveness (Dell'Acqua et al., 2023).

This study aims to contribute to this debate by examining how social artificial agents can influence team performance in complex, cognitively demanding, close-ended environments, using objective performance measures such as completion time and error count, but also more subtle measures such as conversational styles. To this end, we utilize the unique setting of a real-life escape game, as proposed by Englmaier et al., 2023, to study non-routine analytical and interpersonal tasks. In this experimental setup, human teams are challenged with a series of complex problems, such as escaping a room within a time limit, which requires both high cognitive engagement and interactive collaboration. The tasks normally involve finding hidden cues, using objects creatively, and generating innovative solutions to solve quests before time expires. Like real-world non-routine team tasks, these challenges demand diverse perspectives and substantial synergy among team members to achieve success.

Our study builds on this foundation by examining how different embodiments of AI agents - ranging from a simple computer box to an avatar and a humanoid robot - can influence team dynamics and productivity in non-routine analytical tasks. We form teams of three players to complete a real-life ”home-made” escape room, varying team composition across conditions: three humans; two humans and a ’box’; two humans with an avatar; and two humans with the hyper-realistic humanoid robot ABEL from the Centre E. Piaggio (University of Pisa). The escape room involves four distinct challenges, each testing a different skill: linguistic skills, logical and mathematical abilities, knowledge of history and geography, and logical reasoning.

Indeed, the perceived appropriateness of AI agents for different tasks plays a crucial role in shaping human responses. The embodiment and functionality of AI agents also affect how humans interact with them. Research indicates that humans are less receptive to AI agents in tasks perceived as inherently social or ”human”, while they are more accepting of AI involvement in analytical tasks (Chugunova et al., 2022). Research on human-robot teams also indicates that the nature of human-robot collaboration greatly influences teamwork dynamics. For instance, robots perceived as in-group members are more readily accepted and anthropomorphized, which enhances collaboration (Sebo et al., 2020; Fraune, 2020). Furthermore, a robot’s appearance and its ability to express emotions critically shape human behavior and group cohesion (Traeger et al., 2020). Factors such as a robot’s appearance and the prior experience of human team members also significantly affect how well robots are integrated into teams (Destephe et al., 2015).

To understand what constitutes optimal performance within this context, we rely not only on traditional performance metrics, such as the time taken to escape the room and the number of errors made while solving different games, but also on analyzing the linguistic style of team conversation to see whether the presence of an artificial agent influences speech patterns. This focus on language aligns with recent research that has shown the influence of Large Language Models (LLMs), such as ChatGPT, on human spoken and written communication (Yakura et al., 2024). Studies analyzing academic YouTube videos revealed that following the release of ChatGPT, there were noticeable shifts in word usage, suggesting that humans are increasingly adopting linguistic patterns introduced by AI systems (Yakura et al., 2024). These findings highlight concerns about the potential impact of AI on linguistic diversity and how human communication might be evolving due to interaction with AI (Brinkmann et al., 2023). By examining the language used in collaborative environments, our research extends this literature by exploring how AI presence affects not only task performance but also team communication patterns. In particular, we analyze conversations among teams to observe differences in speech with artificial players of different embodiments, relying on Profiling-UD (Brunato et al., 2020), a Computational Stylometry tool, to investigate changes in communication style and Sentence-BERT (SBERT) (Reimers, 2019) for deeper insights into these linguistic changes.

Additionally, we collect data through post-experiment questionnaires to assess the perceptions of individual team members regarding their experiences and interactions. This data allows us to explore correlations between team performance and various characteristics, including both the quantitative metrics of success and the qualitative dynamics of team collaboration and perception.

2 Methodology

2.1 Experimental Design

Our objective is to understand how the composition of collaborative teams affects performance and team working dynamics in a non-routine analytical tasks. Unlike routine tasks that follow a set of standard procedures or rely on repetitive actions, non-routine analytical tasks typically involve unique situations where predefined solutions are not available. To solve these tasks, critical thinking, problem-solving, creativity, and decision-making skills are required. To capture these aspects, our escape room game is thus designed with 4 distinct challenges, each requiring different skill sets:

- Game 1: Linguistic skills are tested.
- Game 2: Logical and mathematical skills are necessary.
- Game 3: Knowledge in history and geography is required.
- Game 4: Logical reasoning abilities are assessed.

In our experiment, participants are grouped in team of three with the main objective to escape the room (i.e. solving all 4 games) within 25 minutes (1500 seconds). They are recruited online from a pool of over 4,000 students across 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 for a hoodie.

Condition	**No Embodiment**	**Low Embodiment**	**High Embodiment**
Machine	Computer Box	Avatar	Abel
No Machine	Only Humans		

Table 1: Experimental condition overview

Across conditions we vary the type of team (all-human teams versus mixed teams of humans and one artificial agent) along with the embodiment of the artificial agent. In particular, we use three different types of artificial agents (see Table 1), 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. ABEL represents the highest level of embodiment among the artificial agents, designed to closely mimic human behavior and appearance. This agent is capable of more nuanced interactions due to its advanced physical and behavioral capabilities, creating a more lifelike and immersive experience for participants [2]

Each participant wears an individual microphone to communicate with the artificial agent, and amongst each other. Although the artificial agent embodiment vary across conditions, they all respond in a similar way. 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

[1] For more information and a demo, see https://github.com/phuselab/openFACS?tab=readme-ov-file

[2] For more information, see https://forelab.unipi.it/technologies/abel-new-generation-hyper-realistic-humanoid-robot.

server that uses Whisper, an advanced artificial intelligence 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 Chat GPT-4, which generates an appropriate response. This response is finally converted into audio using XTTS-v2, an advanced Text-to-Speech (TTS) system.[4] The generated audio is then played back to the player. The LLMs models are integrated into our Otree code borrowing some functionalities from the toolkit developed by Engel et al., 2024.

2.2 Conversational Styles

To study conversational styles, 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 the dialogue structure across all conditions, our analysis specifically focuses on scenarios involving an artificial agent. This focus is necessary because the dialogues in the all-human condition (three humans) are inherently different due to the absence of artificial agents, resulting in a distinct conversational dynamic that is not directly comparable to the conditions with artificial agents. In the human-only condition, participants engage in natural conversations with each other, whereas in the conditions with artificial agents, participants interact in a question-and-answer format with the agent. Therefore, while we can assess how each mixed condition differs from each other, it is not meaningful to directly compare the dialogue dynamics within human-only teams to those within each condition involving an artificial agent in isolation.

To analyze conversational styles in the mixed-team conditions, we utilize the following methodologies:

- Profiling-UD: it is a web-based tool conceived to linguistically profile multilingual texts by relying on the Universal Dependencies (UD) formalism (De Marneffe et al., 2021), a *de facto* standard schema for morpho-syntactic and

[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.

syntactic annotation of corpora, based on the dependency syntactic representation paradigm. It employs Computational Stylometry techniques based on Linguistic Profiling, a methodology originally developed for authorship recognition, which detects and quantifies differences and similarities across texts representing distinct language varieties by analyzing the distribution of numerous linguistic features (Halteren, 2004). As discussed in Section 4.1, 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 (Sentence-BERT) Analysis: This method uses a transformer-based model to assess the semantic similarity between sentences within dialogues (Reimers, 2019). 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 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 an artificial agents need to have to enhance team communication and overall performance in non-routine, analytical tasks.

3 Experimental Evidence

In total, 179 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 four experimental conditions: three humans only (20 sessions), two humans with a computer (*Box*) (16 sessions), two humans with *ABEL* (18 sessions), and two humans with an *Avatar* (23 sessions). We preregistered our analysis plan https://osf.io/g2j7h. Although our study lacks statistical power concerning performance measures and we consider the evidence in this case to be preliminary and exploratory, it provides rich data to study conversational styles (see further below).

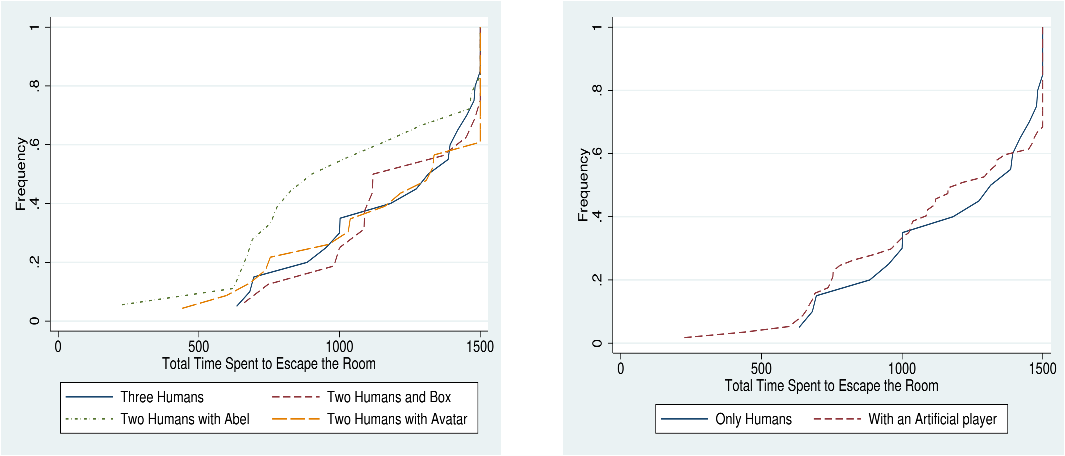
3.1 Results: Team Performance

Table 2 provides an analysis of the proportion of successful teams escaping from the room within the time limit under each experimental condition. The first thing to notice is that for *Human teams* the share of successful all-human teams is notably higher (85%) than for teams with the *Box* (69%), with *Avatar* (57%) and with *Abel* (78%). However, although these results are economically important (on average, the proportion of teams with an artificial player which are successful is 67%) the difference is statistically significant only in contrast with team with Avatar (p-value=0.04). This difference is close to the minimum detectable change in the experimental group proportion, as derived from a power analysis with 80% power. This finding indicates that teams with the Avatar perform significantly worse than all-human teams in successfully escaping the room, and slight worse than teams with other artificial players.

To get a more comprehensive view of each team performance, figure 1 shows the empirical cumulative distribution of finishing times across conditions, with the time limit of 1500 seconds. In addition to highlighting the different proportions of success described above, this figure suggests that, although human-only teams are more successful on average in completing the task, the best-performing groups (i.e., those that escaped the room in the shortest amount of time) are mixed teams with either the avatar or ABEL. This indicates that embodiment matters. In particular, the time taken to escape the room was generally lower for teams with ABEL, even in the central part of the distribution.

Condition 1 vs 2	**Prop 1**	**N1**	**Prop 2**	**N2**	**Difference**	**p-value**
Human vs Box	0.85	20	0.69	20	0.16	0.24
Humans vs Avatar	0.85	20	0.57	23	0.28	0.04
Humans vs Abel	0.85	20	0.78	18	0.07	0.57
Box vs Abel	0.69	16	0.78	18	-0.09	0.55
Box vs Abel	0.69	16	0.57	23	0.12	0.44
Avatar vs Abel	0.78	18	0.57	23	0.21	0.15
Humans vs Artificial	0.85	20	0.67	57	-0.18	0.12

Table 2: Successfully Escaping the room: Proportion Test Results Across Conditions



(a) Human-only teams

(b) Team w/o an artificial player

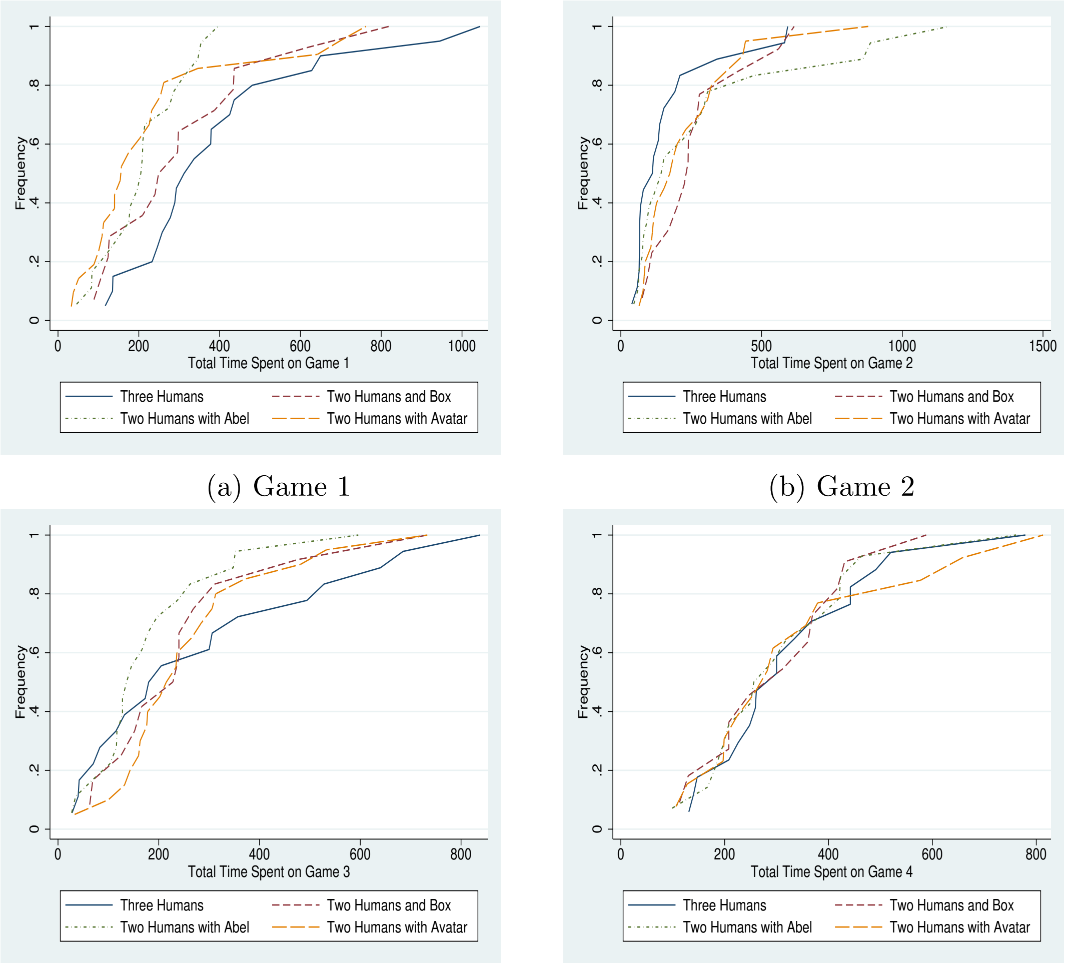
Figure 1: Total time spent to escape the room by condition

To get a more comprehensive view of the performance, we further report the empirical distribution across the four different games and conditions in Figure 2. We can observe that the major differences are seen in Game 1, where having artificial players seems to play a role in achieving higher performance. This effect could be due to the nature of the game itself, as well as the absence of any selection effect, since both successful and unsuccessful teams are present. In contrast, in the fourth and final game, where only high-performing groups remain, and where the role of artificial players is limited (due to the mastermind-type nature of the game), we do not observe any differences across conditions.

	Anderson-Darling		Kolmogorov Smirnov		Mean Time	
	Stat	P-value	Stat	P-value	Humans	Artificial
Total	53.21	0.60	0.14	0.82	1211.45	1153.02
Game 1	79.10	0.03	0.45	0.00	400.05	248.92
Game 2	78.38	0.06	0.33	0.08	171.28	268.04
Game 3	75.35	0.56	0.20	0.55	289.78	233.40
Game 4	76.73	0.25	0.14	0.94	329.06	322.84

Table 3: Comparing distribution time across game

To reinforce this observation, in Table 3 we present the results comparing the time distributions of humans and artificial agents across four games using the Anderson-Darling and Kolmogorov-Smirnov tests, along with the average time spent by each group. For the total comparison across all games, both the Anderson-Darling (Stat = 53.21, p = 0.60) and Kolmogorov-Smirnov (Stat = 0.14, p = 0.82) tests show no significant difference between the distributions of the two groups, suggesting that overall, the time distributions are similar. The mean times



(a) Game 1

(b) Game 2

(c) Game 3

(d) Game 4

Figure 2: Time spent for each game by condition

(Humans = 1211.45, Artificial = 1153.02) further indicate that the two groups took comparable amounts of time overall.[5]

However, looking at the individual games, Game 1 again stands out, with both the Anderson-Darling (p = 0.03) and Kolmogorov-Smirnov (p = 0.00) tests indicating significant differences between humans and artificial agents, where humans took significantly more time (400.05 vs. 248.92). In Game 2, the p-values (0.06 and 0.08) suggest a marginally non-significant difference in the time distributions, though the mean times show humans being faster. Games 3 and 4 show no significant differences (all p-values well above 0.05), with both groups performing

[5] Both the KS and AD tests are based on the cumulative probability distribution of data. They both calculate the distance between distributions at each point along the scale. The AD test is more powerful, particularly due to its sensitivity to the shape and scale of a distribution and differences in the tails of distributions. Additionally, the AD test requires less data than the KS test to achieve sufficient statistical power. See e.g. Engmann et al., 2011; Baumgartner et al., 2023.

similarly in terms of time spent. Thus, the main difference in performance appears in Game 1, while the other games exhibit closer performance between the two groups. This result further suggests that it is in the first game, where no selection effects occur among teams, that having artificial players seems to play a role in achieving higher performance.

Table 4 reports the total number of errors across conditions. The results indicate a key difference in the performance of teams depending on their composition, particularly when comparing human-only teams to mixed teams involving an artificial agent (Box, Abel, and Avatar). On average, human-only teams made significantly more errors compared to teams that included any artificial agent (18.67 vs. 10.14). This difference is both economically (8.53) and statistically significant (p-value = 0.02). This suggests that the inclusion of an artificial agent, regardless of its type, tends to reduce the number of errors made by the team. Looking at the average number of errors per completed game (not-reported) does not change the conclusions. Further examination of the comparisons between human-only teams and each specific type of artificial agent (Box, Abel, and Avatar), as well as each type of game, shows that while human teams generally made more errors, these differences are not statistically significant when considered individually (p-values ranging from 0.10 to 0.18). Statistically, this suggests that there may be no significant difference in error rates due to limited statistical power.

However, a clear difference emerges in Game 3: in this case, not only do we observe the advantage of having an artificial player on the team (the difference being 3.39, p-value = 0.02), but there is also a clear advantage of playing with Abel, the artificial agent with the highest level of embodiment (difference = 5.61, p-value = 0.01).

There are also indications of a non-linear relationship, with the highest error rates observed at medium levels of embodiment. This implies that although artificial agents might generally offer an advantage in reducing errors, the specific type of artificial agent — whether a simple Box, an Avatar, or the hyper-realistic humanoid Abel— may significantly impact the team's error rate.

In Table 11 in the Appendix, we further compare the number of errors by categorizing the groups into ”Success” and ”Failure,” indicating whether the teams succeeded in completing the game or not. Overall, the results confirm that human-only teams tend to make more errors than teams with an artificial agent. Notably, human-only teams, even when successful, generally commit more errors. Furthermore, our findings suggest a non-linear relationship between the number of errors and the embodiment of the artificial agents; however, this observation is limited

Table 4: T-test Results for Errors Across Games by Experiment Type

	Mean 1	**N1**	**Mean 2**	**N2**	**Diff**	**p-value**
Total Errors						
Humans vs Box	18.67	18	7.64	11	11.03	0.07
Humans vs Abel	18.67	18	8.40	15	10.27	0.06
Humans vs Avatar	18.67	18	13.29	17	5.37	0.31
Box vs Abel	7.64	11	8.40	15	-0.76	0.78
Box vs Avatar	7.64	11	13.29	17	-5.66	0.11
Abel vs Avatar	8.40	15	13.29	17	-4.89	0.14
Humans vs Artificial	18.67	18	10.14	43	8.53	0.02
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
Humans vs Artificial	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
Humans vs Artificial	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
Humans vs Artificial	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
Humans vs Artificial	2.72	18	2.07	43	0.65	0.51

by the study's statistical power, which affects the reliability of these more detailed comparisons.

These findings, combined with the success rates presented above, suggest that while artificial agents generally reduce errors, their presence might also introduce new dynamics that affect a team's overall success. This influence may vary depending on the embodiment of the artificial agent, indicating that different levels of embodiment could impact team dynamics and outcomes in unique ways.

To sum up, we can conclude that human-only teams are, on average, more successful when playing alone, even though they tend to make more mistakes. However, the best-performing groups are those that play alongside an artificial agent with the highest form of embodiment. Moreover, there appears to be a non-linear relationship between team performance and the embodiment of the artificial player.

4 Results: Conversation Dynamics

In the following, we will analyze the dialogue within each team across conditions. Table 5 reports the main summary statistics regarding the average number of messages exchanged per human per session, across conditions and games.

Experiment Type	**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
Total	1.75	0.75	0.99	0.61	4.11

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

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 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 earlier, 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 4.1 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 4.2 we discuss the outcomes of an analysis focused on the content of the messages exchanged by humans with the three experimental conditions.

4.1 Analysis of Communication Style

As previously mentioned, for the analysis of the communication style of the messages we relied on Profiling-UD (Brunato et al., 2020), 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 (Straka et al., 2016), a state-of-the-art pipeline available for nearly all languages in the Universal Dependencies (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 7, 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 modeled, 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

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
Distibution 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

Table 6: Linguistic Features extracted by Profiling-UD categorized into nine groups of linguistic phenomena.

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 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.

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

Table 7: Average values of raw text features extracted from messages exchanged between humans and the three artificial agents.

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-

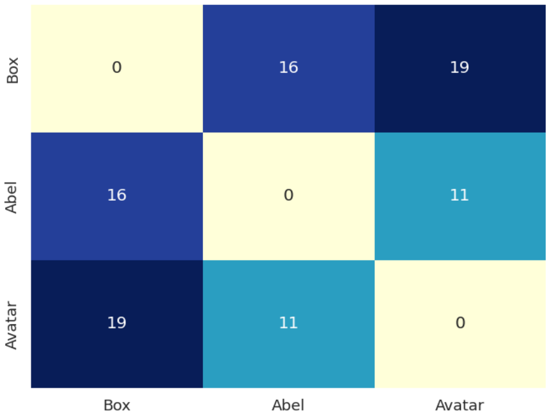


Figure 3: The number of statistically different features extracted by Profiling-UD across three experimental conditions, excluding the human-only condition.

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 3. 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 embodiments of the artificial agents are higher and more similar to each other, such as Avatar and Abel.

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 (Wendt, 1972), which ranges from -1 to $+1$. 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, 9 and 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.

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

Table 8: Linguistic features that are statistically ($p < 0.05$) different between Abel and Box conditions. Mean values and standard deviations (±) are reported. Features in each group are ordered by absolute decreasing rank-biserial correlation value (r). $\star$ marks highly statistically significant features ($p < 0.01$).

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 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*.

Group	**Feature**	**Box**	**Avatar**	***r***
Morphosyntax	% adpositions	7.14±7.33	8.99±7.35	0.15
	% adjectives	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

Table 9: Linguistic features that are statistically ($p < 0.05$) different between Avatar and Box conditions. 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 ($p < 0.01$).

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.

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

Table 10: Linguistic features that are statistically ($p < 0.05$) different between Abel and Avatar conditions. Mean values and standard deviations (±) are reported. Features in each group are ordered by absolute decreasing rank-biserial correlation value (r). $\star$ marks highly statistically significant features ($p < 0.01$).

4.2 Sentence SBERT

In this analysis, we compute the SBERT (Sentence-BERT) score for each dialogue that occurred across different conditions involving an artificial player, complementing the linguistic profiling analysis described above. Since the dialogue structure in the human-only condition differs significantly, as stated before, we will focus on the semantic alignment of meaning during interactions with these artificial agents.

Specifically, we measure the semantic similarity of each sentence by comparing it to corresponding sentences in other artificial conditions and then taking the maximum similarity score. The average of these maximum cosine similarities is calculated to provide an overall measure of semantic alignment for each session under each condition. As mentioned earlier, cosine similarity ranges from -1 (maximum dissimilarity) to 1 (maximum similarity).

Figure 4 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 —using a t-test - 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 (though still far from 1). The results are robust across individual games, with the largest similarities appearing in Game 3 and the smallest in Game 4. 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.

In line with the results presented above, these findings suggest that the type of artificial player— whether Box, Abel, or Avatar— affects the team's overall dialogue. Indeed, we observe the lowest cosine similarities in Game 1, where the highest performance difference are also observed. Moreover, the results indicate that more embodied artificial players, like Abel, may interact with human team members in a way that generates distinct conversational patterns compared to simpler representations like Box and Avatar.

5 Conclusions

In this study, using a real-life escape room scenario, we investigated the influence of an artificial agent on team performance and communication style in non-routine analytical tasks. Participants worked in teams composed either entirely of humans or with an artificial agent that differed in its embodiment, ranging from a simple box to a highly sophisticated humanoid robot, Abel. Across four distinct subgames, each testing different skill sets, the study evaluated team performance in

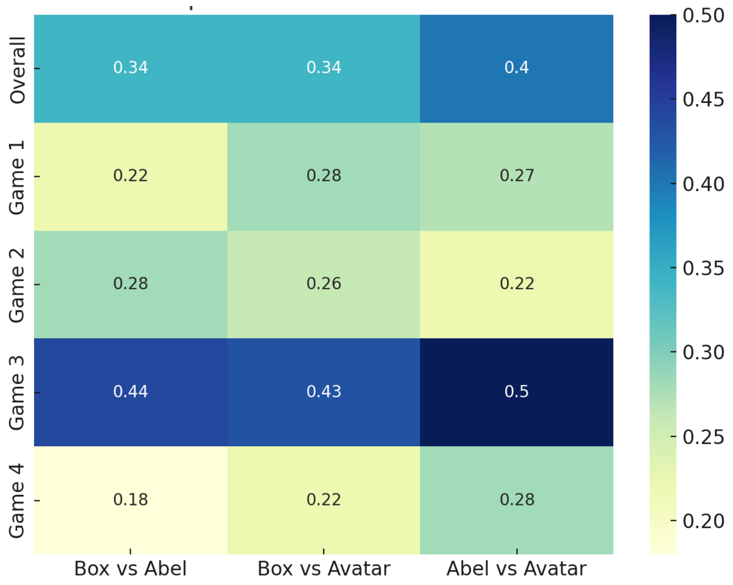


Figure 4: SBERT score average similarity

terms of success rates, time taken, and the number of errors made. Additionally, we analyzed the communication patterns of the participants to assess how interactions with different artificial embodiments affected dialogue.

The results indicate that human-only teams had a higher success rate in escaping the room, with 85% completing the task, compared to (an average of) 67% for teams that included an artificial agent. However, teams with artificial agents, particularly the more sophisticated embodiment like Abel, made significantly fewer errors across the games. Although the overall time to complete the tasks did not show a significant difference between human-only and mixed teams, Abel's presence was associated with faster performance in certain games, suggesting that higher levels of embodiment might improve efficiency in some scenarios. Overall, these findings highlight a trade-off between success rates and error minimization when collaborating with artificial agents.

In terms of communication, the analysis revealed that interacting with more embodied agents, such as Abel, led to more complex and natural dialogue compared to simpler agents like the box or avatar. Human participants adapted their communication styles depending on the agent's level of embodiment, with more human-like agents fostering richer and more conversational interactions. 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. Although still an exploratory study, our results are strongly in line with other research that suggests interacting with artificial players (i.e., LLM models) may affect human communication styles beyond text and, in turn, influence human cultural evolution.

Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this work the authors used Chatgpt4 in order to improve readbility 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.

References

Baumgartner, Daniel and John Kolassa (2023). “Power considerations for Kolmogorov–Smirnov and Anderson–Darling two-sample tests”. In: *Communications in Statistics-Simulation and Computation* 52.7, pp. 3137–3145.

Brinkmann, Levin et al. (2023). “Machine culture”. In: *Nature Human Behaviour* 7.11, pp. 1855–1868.

Brunato, Dominique et al. (2020). “Profiling-ud: a tool for linguistic profiling of texts”. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 7145–7151.

Chugunova, M. and D. Sele (2022). “We and It: An interdisciplinary review of the experimental evidence on how humans interact with machines”. In: *Journal of Behavioral and Experimental Economics.*

Corgnet, Brice, Roberto Hernán-González, and Ricardo Mateo (2023). “Peer effects in an automated world”. In: *Labour Economics* 85, p. 102455.

De Marneffe, Marie-Catherine et al. (2021). “Universal dependencies”. In: *Computational linguistics* 47.2, pp. 255–308.

Dell’Acqua, Fabrizio et al. (2023). “Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality”. In: *Harvard Business School Technology & Operations Mgt. Unit Working Paper* 24-013.

Destephe, Matthieu et al. (2015). “Walking in the uncanny valley: Importance of the attractiveness on the acceptance of a robot as a working partner”. In: *Frontiers in psychology* 6, p. 204.

Engel, Christoph, Max RP Grossmann, and Axel Ockenfels (2024). “Integrating machine behavior into human subject experiments: A user-friendly toolkit and illustrations”. In: *MPI Collective Goods Discussion Paper* 2024/1.

Englmaier, Florian et al. (2023). “The effect of incentives in non-routine analytical teams tasks-evidence from a field experiment”. In: *Journal of Political Economy.*

Engmann, Sonja and Denis Cousineau (2011). “Comparing distributions: the two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test.” In: *Journal of applied quantitative methods* 6.3.

Fraune, Marlena R (2020). “Our robots, our team: Robot anthropomorphism moderates group effects in human–robot teams”. In: *Frontiers in psychology* 11, p. 1275.

Gombolay, Matthew C et al. (2015). “Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams”. In: *Autonomous Robots* 39, pp. 293–312.

Halteren, Hans van (2004). “Linguistic profiling for author recognition and verification”. In: *Proc. of ACL*, pp. 200–207.

Johnson, Matthew et al. (2012). “Autonomy and interdependence in human-agent-robot teams”. In: *IEEE Intelligent Systems* 27.2, pp. 43–51.

Li, Ning, Huaikang Zhou, and Kris Mikel-Hong (2024). “Generative AI Enhances Team Performance and Reduces Need for Traditional Teams”. In: *arXiv preprint arXiv:2405.17924*.

Otis, Nicholas et al. (2023). “The uneven impact of generative AI on entrepreneurial performance”. In: *Available at SSRN 4671369*.

Reimers, N (2019). “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *arXiv preprint arXiv:1908.10084*.

Sebo, Sarah et al. (2020). “Robots in groups and teams: a literature review”. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2, pp. 1–36.

Straka, Milan, Jan Hajič, and Jana Straková (May 2016). “UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Ed. by Nicoletta Calzolari et al. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 4290–4297. URL: https://aclanthology.org/L16-1680.

Traeger, Margaret L et al. (2020). “Vulnerable robots positively shape human conversational dynamics in a human–robot team”. In: *Proceedings of the National Academy of Sciences* 117.12, pp. 6370–6375.

Wendt, Hans W (1972). “Dealing with a common problem in social science: A simplified rank-biserial coefficient of correlation based on the statistic.” In: *European J. of Social Psychology*.

Yakura, Hiromu et al. (2024). “Empirical evidence of Large Language Model’s influence on human spoken communication”. In: *arXiv preprint arXiv:2409.01754*.

A Additional tables

	Success						Failure					
	Mean 1	**N1**	**Mean 2**	**N2**	**Diff**	**p-value**	**Mean 1**	**N1**	**Mean 2**	**N2**	**Diff**	**p-value**
Game 1 Errors												
Humans vs Box	6.53	17	1.09	17	5.44	0.28	15.33	3	26.80	5	-11.47	0.57
Humans vs Abel	6.53	17	3.64	17	2.89	0.52	15.33	3	1.75	4	13.58	0.16
Humans vs Avatar	6.53	17	1.85	17	4.68	0.31	15.33	3	28.00	10	-12.67	0.69
Box vs Abel	1.09	11	3.64	11	-2.55	0.05	26.80	5	1.75	4	25.05	0.15
Box vs Avatar	1.09	11	1.85	11	-0.76	0.29	26.80	5	28.00	10	-1.20	0.96
Abel vs Avatar	3.64	14	1.85	14	1.80	0.14	1.75	4	28.00	10	-26.25	0.35
only humans	6.53	17	2.29	17	4.24	0.12	15.33	3	22.16	19	-6.82	0.78
Game 2 Errors												
Humans vs Box	2.06	17	2.55	17	-0.49	0.76	6.33	3	6.00	5	0.33	0.95
Humans vs Abel	2.06	17	2.14	17	-0.08	0.96	6.33	3	4.75	4	1.58	0.76
Humans vs Avatar	2.06	17	1.46	17	0.60	0.69	6.33	3	3.00	10	3.33	0.32
Box vs Abel	2.55	11	2.14	11	0.40	0.80	6.00	5	4.75	4	1.25	0.75
Box vs Avatar	2.55	11	1.46	11	1.08	0.49	6.00	5	3.00	10	3.00	0.25
Abel vs Avatar	2.14	14	1.46	14	0.68	0.65	4.75	4	3.00	10	1.75	0.51
only humans	2.06	17	2.03	17	0.03	0.98	6.33	3	4.16	19	2.18	0.50
Game 3 Errors												
Humans vs Box	6.76	17	2.64	17	4.13	0.16	6.67	3	1.80	5	4.87	0.38
Humans vs Abel	6.76	17	1.57	17	5.19	0.04	6.67	3	3.00	4	3.67	0.55
Humans vs Avatar	6.76	17	5.62	17	1.15	0.71	6.67	3	2.80	10	3.87	0.32
Box vs Abel	2.64	11	1.57	11	1.06	0.41	1.80	5	3.00	4	-1.20	0.52
Box vs Avatar	2.64	11	5.62	11	-2.98	0.28	1.80	5	2.80	10	-1.00	0.55
Abel vs Avatar	1.57	14	5.62	14	-4.04	0.08	3.00	4	2.80	10	0.20	0.91
only humans	6.76	17	3.26	17	3.50	0.08	6.67	3	2.58	19	4.09	0.16
Game 4 Errors												
Humans vs Box	2.29	17	1.36	17	0.93	0.38	3.33	3	0.00	5	3.33	0.22
Humans vs Abel	2.29	17	1.00	17	1.29	0.13	3.33	3	1.50	4	1.83	0.60
Humans vs Avatar	2.29	17	0.92	17	1.37	0.11	3.33	3	4.20	10	-0.87	0.84
Box vs Abel	1.36	11	1.00	11	0.36	0.63	0.00	5	1.50	4	-1.50	0.29
Box vs Avatar	1.36	11	0.92	11	0.44	0.56	0.00	5	4.20	10	-4.20	0.17
Abel vs Avatar	1.00	14	0.92	14	0.08	0.87	1.50	4	4.20	10	-2.70	0.44
only humans	2.29	17	1.08	17	1.22	0.05	3.33	3	2.53	19	0.81	0.80

Table 11: T-test Results for Errors Across Games by Success and Experiment Type