Physiologically-Informed Predictability of a Teammate's Future Actions Forecasts Team Performance

Yinuo Qin[1*], Richard T. Lee[1,2], Weijia Zhang[1], Xiaoxiao Sun[1], Paul Sajda[1,2,3*]

[1]Department of Biomedical Engineering, Columbia University, New York, NY, USA.
[2]Department of Electrical Engineering, Columbia University, New York, NY, USA.
[3]Department of Radiology, Columbia University, New York, NY, USA.

*Corresponding author(s). E-mail(s): yinuo.qin@columbia.edu; psajda@columbia.edu;
Contributing authors: rtl2118@columbia.edu; wz2540@columbia.edu; xiaoxiao.sun@columbia.edu;

Abstract

In collaborative environments, a deep understanding of multi-human teaming dynamics is essential for optimizing performance. However, the relationship between individuals' behavioral and physiological markers and their combined influence on overall team performance remains poorly understood. To explore this, we designed a triadic human collaborative sensorimotor task in virtual reality (VR) and introduced a novel predictability metric to examine team dynamics and performance. Our findings reveal a strong connection between team performance and the predictability of a team member's future actions based on other team members' behavioral and physiological data. Contrary to conventional wisdom that high-performing teams are highly synchronized, our results suggest that physiological and behavioral synchronizations among team members have a limited correlation with team performance. These insights provide a new quantitative framework for understanding multi-human teaming, paving the way for deeper insights into team dynamics and performance.

1 Introduction

Teamwork is a critical form of human interaction, productivity, and survival. From world-championship sports teams to intimate working groups, from ancient tribal rituals to modern urban planning, teaming has consistently been a critical and innate element of human behavior. Without teaming, our society would likely look very different, lack rich and diverse cultures, lack marvels of engineering construction, and have limited groundbreaking scientific advancements. Studying the fundamental mechanisms behind human teaming is essential to understanding and improving collective human intelligence.

Games and collaborative tasks have been used as major platforms to study multi-human teaming. From role-playing to battle arena games, many previous studies have shown that multiplayer online games have great potential for studying team dynamics (6), leadership in multi-human teaming (19; 31), and individual behavior within teams (36). However, it is still unclear whether the insights gained from simple game-based studies can be generalized to more complex, high-stakes team interactions and team performance.

In addition to computer games, real-world scenarios, such as simulated hospitals with surgical teams and teams in manufacturing companies, have been used to study team performance and effectiveness (28; 11). Most of these previous studies have used qualitative methods such as interviews (11), questionnaires (11; 30), and surveys (28; 42). While these qualitative studies can help us gain insights into how some task-related factors can impact team performance, these methods are prone to bias due to subjective reporting and are often difficult to reproduce. Therefore, additional consideration of quantitative evaluation metrics to understand team performance remains essential but largely unexplored.

With the development of virtual reality (VR), more environmentally controlled team-based studies have been conducted (43; 15; 27; 46). VR provides an immersive experience while reducing external distractions. The virtual environment also has the potential to provide realistic simulations with well-controlled delivery and simultaneous recording of events and interactions. However, few VR experiments have involved real-time synchronization and multi-modal data collection of multi-person teams. Most team-based VR experiments are conducted with a single human performing collaborative tasks with other simulated computer agents instead of working in the simulation with other people (15; 27). These experiments limit the possibility of studying multi-human teaming.

Through studying human teaming in various tasks, previous research has highlighted that physiological synchrony among team members is positively correlated with team performance (18; 8; 13; 23). Conversely, other studies have suggested a negative correlation between behavioral synchrony and team performance (2; 45). The preceding literature lacks studies that comprehensively correlate performance with both behavioral and physiological synchrony in complex teaming tasks. The interpretation of such correlations of team performance with behavioral synchrony and physiological synchrony remains unclear and incomplete. Therefore, we hypothesized that a comprehensive understanding of the balance between physiological and behavioral synchrony

is critical for enhancing team performance, especially in tasks that demand high levels of cooperation, coordination, and collaboration.

In this work, we developed a novel framework to study multi-human teaming in a VR environment by quantitatively analyzing multi-modal physiological and behavioral data from all team members. We constructed an immersive sensorimotor task requiring three participants to collaboratively navigate a spacecraft, capturing multi-modal data from all participants. To identify potential biomarkers of team performance, we employed two computational approaches: inter-subject correlation (ISC) and predictability. ISC, traditionally linked to team performance metrics (41; 33), was found to correlate with team performance only under specific measurements in our complex collaborative task. To address these limitations, we proposed a predictability approach, using a deep learning model to forecast one team member's remote controller actions based on their teammates' physiological and behavioral data. This predictive model revealed a significant correlation between the predictability of team members' actions and team performance, suggesting that predictability can serve as a robust biomarker for understanding and enhancing team dynamics in collaborative tasks.

2 Results

To test the correlation between team performance and physiological and behavioral synchrony among team members, we designed a multi-human team-based virtual reality (VR) task that we refer to as the **A**pollo **D**istributed **C**ontrol **T**ask (ADCT). Our task is inspired by the renowned Apollo 13 reentry mission and its extended cinematic story (34; 1). The Apollo 13 mission is considered one of history's most "successful failures" in that three astronauts exhibited extraordinary teamwork while operating different controls of a spacecraft collaboratively to safely navigate back to Earth after an oxygen tank exploded. The ADCT is a team-based version of a boundary avoidance task (BAT) (12), which requires substantial attention and regulation of arousal of each individual in the team. The ADCT has the following features built into its design and construction: 1) it is a challenging enough cooperative and collaborative task to trigger complex team dynamics; 2) the experiment was conducted repetitively with a consistent group of participants; 3) the task state and behavior of subjects are synchronized in real-time with simultaneously recorded multi-modal physiological signals; 4) team performance is quantitatively assessed by evaluating the contributions of all team members, where local performance is measured in relation to short-term goals, and global performance encompasses high-level planning tragedies. Fig. 1 summarizes the ADCT.

Specifically, the ADCT is performed by a triad team in VR (Fig.1 a). Each team member, as a co-pilot, has partial observation of the exterior space environment through uniquely positioned spacecraft windows, each with different viewing points. Each co-pilot controls a single degree of freedom of the spacecraft's movement, such as yaw, pitch, or thrust (Fig.1 b). The team's goal is to collaboratively navigate the spacecraft back to Earth by following a pre-defined reentry path. The transparent red rings mark the boundary of the path, and the team must reach Earth within a limited time. Therefore, failing to pass all rings with sufficient speed results in trial failure.

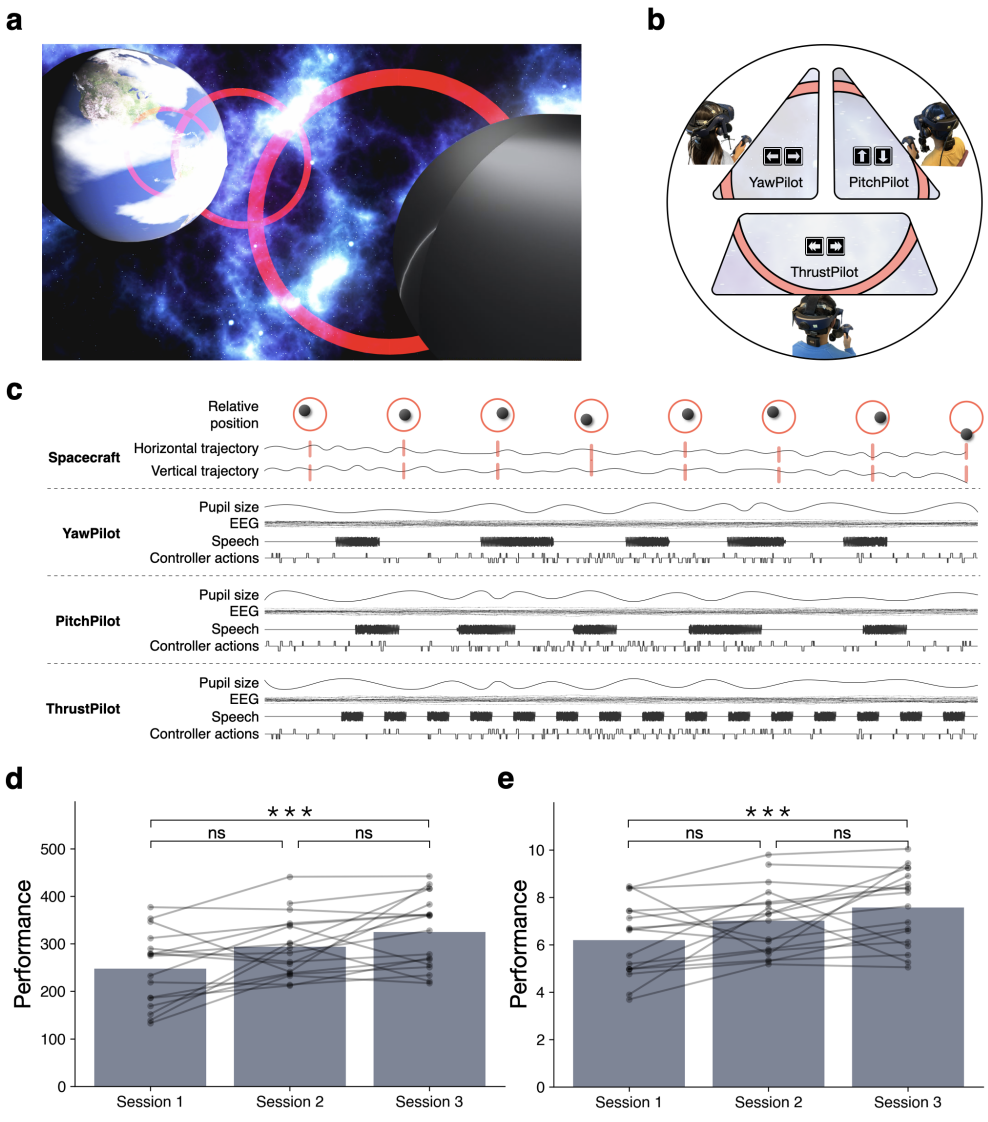

Fig. 1 ADCT environment and team performance. **a**, An illustration of ADCT virtual environment. The team's goal is to control the spacecraft, passing all red rings and arrive back to Earth within the specified time limit. **b**, The view of three co-pilots with respect to a ring obstacle and the degree of freedom controlled by each role. The three co-pilots are YawPilot, PitchPilot, and ThrustPilot. Each participant was equipped with a VR headset, a microphone, a remote controller, and an EEG headset. **c**, Illustration of data modalities collected from all co-pilots. The red bars on the spacecraft's horizontal and vertical trajectories represent the relative location of ring obstacles. The uppermost section illustrates the cross-section of a spacecraft's position with respect to a ring. **d-e**, Team performance across three experimental sessions. The performance is measured by the number of rings passed. Each dot indicates one team (N = 17). Bars indicate the average across teams. **d**, Total number of rings passed by each team in each session. **e**, Averaged the number of rings passed by each team in each trial in three sessions. Asterisks indicate statistically significant differences, defined as ns, not significant, $***P < 0.001$.

Teams are monetarily incentivized to complete as many trials successfully as possible. If they cannot return to Earth in time, they must navigate the entry path to get as close to Earth as possible.

While the teams performed the ADCT, we simultaneously collected electroencephalography (EEG), pupillometry, eye gaze, speech, and remote controller inputs from all participants (Fig.1 c). Each team participated in three experimental sessions. The roles of participants were randomly assigned for each session, but the team members remained the same across all sessions. Each experimental session included 45 trials, each consisting of 15 rings. Team performance was quantitatively evaluated by the team's total number of ring obstacles successfully navigated.

2.1 Team Performance Improves Across Experimental Sessions

We first analyzed performance dynamics across three experimental sessions to investigate how physiological and behavioral synchrony among team members relates to team performance. As shown in Fig.1 d, the total number of rings passed by each team increased monotonically over the experimental sessions, indicating a steady improvement in overall team performance. Repeated measures analyses of variance (ANOVA) revealed significant performance differences across sessions ($F(2, 32) = 10.88,\ p < 0.001$,). Post-hoc comparisons with Bonferroni correction showed a substantial improvement in performance from Session 1 to Session 3 ($p < 0.001$). Similarly, the averaged trial performance also improved significantly over time (Fig. 1 e, $F(2, 32) = 7.75,\ p < 0.01$). The performance significantly increased from Session 1 to Session 3 ($p < 0.001$). These findings suggest that team performance improved consistently as participants engaged in more task sessions. This steady enhancement highlights the potential for learning and adaptation in team dynamics through repeated collaborative tasks in immersive environments.

2.2 Subjective Ratings and Multi-Modal Inter-Subject Synchrony

After each experimental session, all co-pilots provided subjective ratings of their familiarity with and helpfulness toward other team members (see Post-Task Survey in Post Task Survey for details). Analyzing these ratings allows us to track how familiarity and helpfulness change over time and investigate the potential impact of team members' perceptions on team performance. Surprisingly, the helpfulness rating shows a consistent decrease across the experimental sessions (Fig. 2 a, repeated measures ANOVA, $F(2, 34) = 9.33, p < 0.001$). In contrast to the decreasing helpfulness scores, the average familiarity rating across teams increases monotonically (Fig.2 b, $F(2, 34) = 21.42, p < 0.001$). This pattern suggests that as team members become more familiar with each other, their perceptions of helpfulness may become more critical or nuanced.

Next, we analyzed the dynamics of team synchronization by calculating the inter-subject correlation (ISC) across various data modalities. ISC is a widely recognized metric for evaluating the synchrony among individuals performing identical tasks (20; 9; 8; 32; 21) or collaborative tasks (39; 49). This work analyzed the ISC among

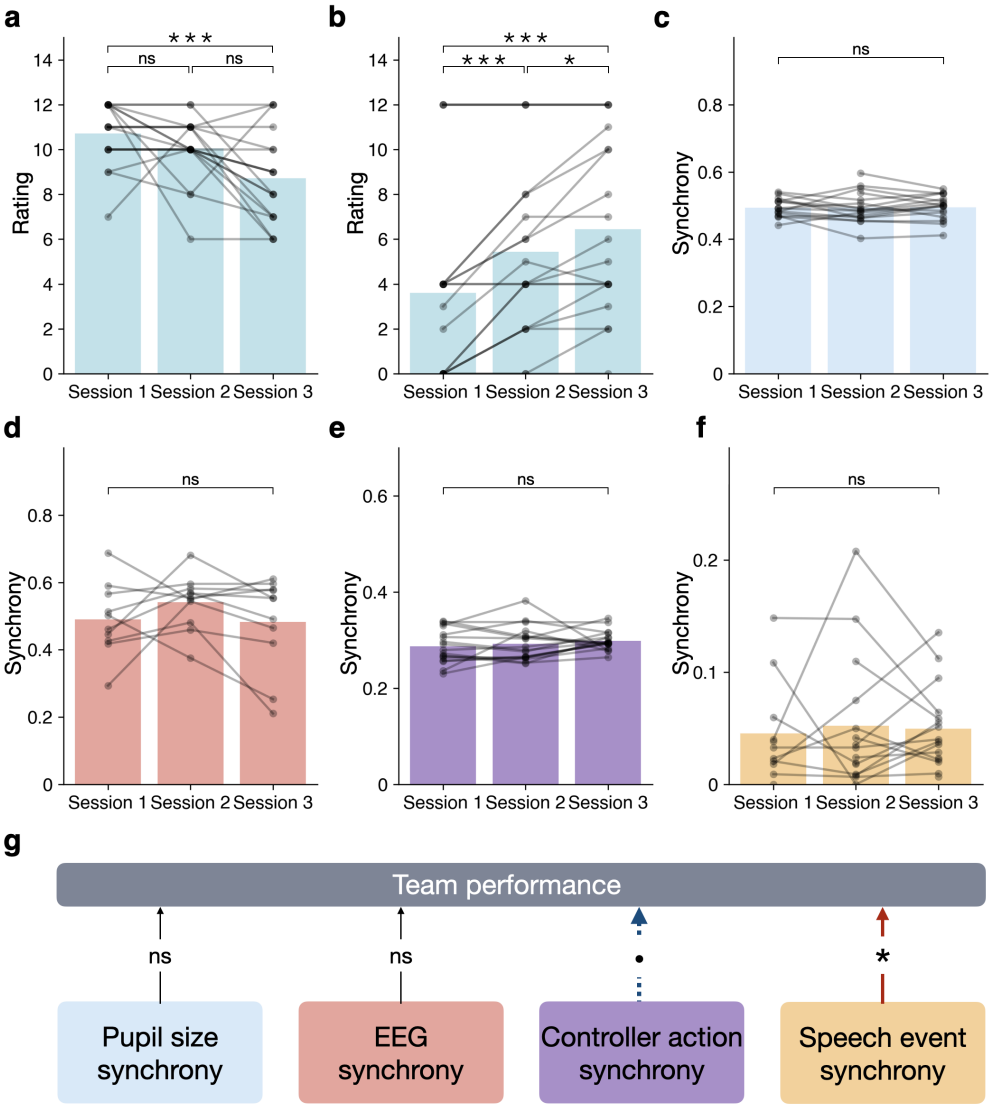

Fig. 2 Subjective rating and multi-modal synchrony. **a-f**, Subjective ratings and synchrony among team members based on different physiological or behavioral data modalities across experimental sessions. Each dot represents one team, and the bars show the average across teams. **a**, Helpfulness rating of team members (N=17). **b**, Familiarity rating of team members (N=17). **c**, Pupil size synchrony among team members (N=14). **d**, EEG synchrony among team members (N=9). **e**, Remote controller action synchrony among team members (N=17). **f**, Speech event synchrony among team members (N=7). **g**, Multi-modal synchrony and its correlation with team performance. Blue arrows indicate negative correlations, while red arrows indicate positive correlations. Asterisks indicate statistically significant differences, defined as ns, not significant, $\cdot P < 0.1, *P < 0.05, ***P < 0.001$.

three co-pilots based on their pupil dynamics, EEG, remote controller inputs, and speech events. As illustrated in Fig.2 c, pupil size synchrony remains relatively stable across sessions ($F(2, 28) = 0.65, p = 0.53$). Interestingly, EEG ISC is maximized in the second experimental session. However, variations in EEG ISC did not achieve statistical significance (Fig.2 d, $F(2, 18) = 1.51, p = 0.25$). Remote controller actions and speech events also remain stable along the three experimental sessions (Fig.2 e, f; remote controller actions synchrony $F(2, 32) = 1.10, p = 0.34$; speech event synchrony $F(2, 14) = 0.23, p = 0.80$). These findings suggest that increasing experimental sessions have a limited impact on synchronizations among team members' behavioral or physiological data.

2.3 Inter-Subject Synchrony and Its Correlation with Team Performance

Inter-subject synchrony (ISC) is often hypothesized to be correlated with team performance. Previous studies have demonstrated a positive relationship between team performance and synchrony in brain and pupil dynamics (41; 8; 33; 48). However, whether synchrony among more than two team members correlates with overall team performance remains unexplored. To address this, we employed generalized linear mixed-effects models (GLMMs) to examine the relationship between inter-subject synchrony across multiple modalities and team performance (see Generalized Linear Mixed-effect Model for details).

Our findings reveal that behavioral synchrony, such as controller action synchrony and speech event synchrony, significantly correlate with team performance (Fig. 2 g). Interestingly, speech event synchrony among team members is positively correlated with team performance, suggesting that verbal coordination enhances high-level task outcomes ($\beta = 1.63, P = 0.039$). In contrast, controller action synchrony is negatively correlated with team performance, possibly reflecting a detrimental effect of over-coordination on individual autonomy in control actions ($\beta = -1.01, P = 0.072$). Physiological synchrony, however, did not show a significant correlation with team performance (pupil size synchrony, $\beta = -0.73, P = 0.328$; EEG synchrony, $\beta = 0.11, P = 0.845$). These results show that behavioral synchrony is a key predictor of team performance in triad teams, highlighting a previously overlooked factor in team performance research.

2.4 Quantifying Team Predictability Using Multi-Modal Physiological and Behavioral Data

A high-performing team consists of members who consistently engage in predictable interactions (3). This predictability results from a deep understanding and harmony within the team, making it easier for team members to anticipate one another's actions and reactions to each other. In this study, we used a multi-head attention model to quantitatively measure how the future actions of a teammate could be predicted from their teammates' physiology and behavior.

First, we epoched multi-modal physiological and behavioral data from 1.5 seconds before each ring-passing event (Fig.3 a). The model received inputs from the

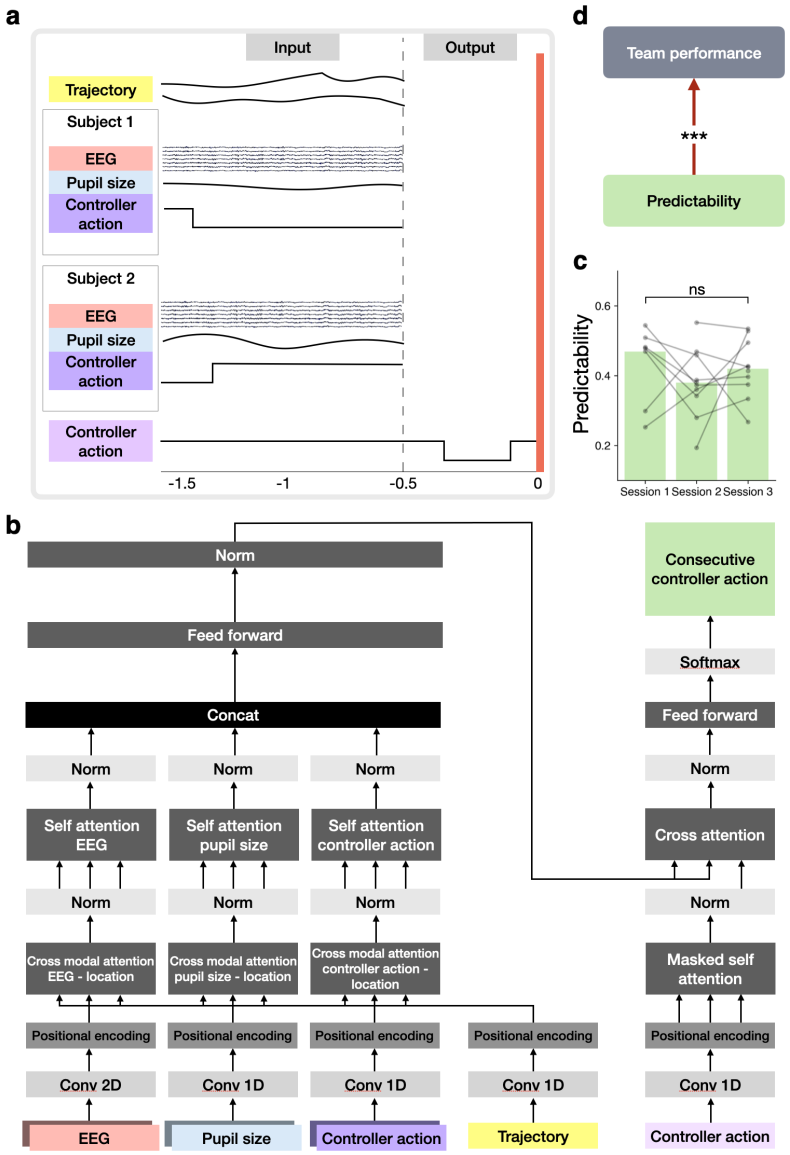

Fig. 3 Predictability of each team member's actions as a biomarker of team performance. **a**, An illustration of a single epoch of the multi-modal data. Each epoch is relative to a ring, and we divided each epoch into input and output for the predictive model. The predicted action of an individual is based on a generative model that uses the behavioral and physiological data of the other two co-pilots. Predictability is evaluated by computing the correlation of the true action of a co-pilot with the model-predicted action. **b**, Multi-head attention modal structure. The cross-modal attention layers take the spacecraft trajectories and physiological or behavioral data. **c** Team predictability across three experimental sessions. Each dot represents one team, and the bars show the average across teams ($n = 10$). ns, not significant. **d**, Correlation between team performance and predictability. The red arrow indicates positive correlations. Asterisks indicate statistically significant differences, defined as $***P < 0.001$.

initial 1 second of this epoch, where each input included the spacecraft's trajectory and the behavioral and physiological data of two co-pilots. The model's output was the generated prediction of the constructive 0.5-second controller action of the third co-pilot (0.5 seconds before passing the ring). On average, co-pilots made about 0.3 remote controller actions in that time period. We evaluated predictability at the team level by averaging the individual predictability scores across the three co-pilots. Since speech event synchrony significantly correlates with team performance ($P < 0.05$), we excluded speech event data from the model input to avoid potential bias. (The supplementary materials include results from a model incorporating speech input for comparison.) By analyzing team predictability, we demonstrate its potential as a biomarker significantly associated with overall team performance.

This model architecture addresses the challenge of integrating multi-modal data with varying temporal and spatial characteristics (Fig.3 b). Including cross-modal attention layers ensures that inter-modal dependencies are captured, particularly when aligning trajectory information across diverse behavioral and physiological data modalities. The self-attention layers for each modality are crucial for extracting meaningful intra-modal patterns, such as EEG synchrony or patterns in pupil size dynamics. By concatenating the outputs of all modalities, the feed-forward network integrates complementary features, creating a unified representation that captures the interactions between physiological, behavioral, and environmental data.

The cross-attention mechanism bridges the fused multi-modal representation with the target modality, facilitating accurate predictions of controller actions. This structure allows the model to leverage the unique contributions of each modality while ensuring robustness in handling noisy data. The design promotes modularity and adaptability, making it suitable for analyzing multi-modal tasks requiring temporal and spatial alignment, such as collaborative team performance or dynamic decision-making tasks.

We hypothesized that the predictability of team members' future controller actions would significantly correlate with team performance. Consequently, we expected that the predictability of each team's actions would change across experimental sessions. As shown in Fig. 3 c, predictability changes slightly as the number of experimental sessions increases ($F(2, 10) = 0.12, P = 0.888$). A detailed analysis of predictability and team performance is provided in the next section.

2.5 Team Action Predictability as a Performance Biomarker

We derive an intriguing finding that predictability serves as an important biomarker for team performance (Fig. 3 d, $\beta = 3.20, P < 0.001$). The positive correlation with team performance suggests that when team members can better anticipate each other's future actions, overall coordination improves, enabling the team to achieve higher-level goals such as passing more rings.

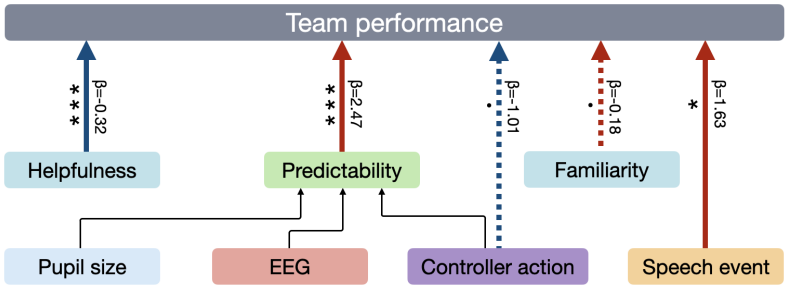

Fig. 4 Overview of the correlation between predictability and team performance. All correlations are the GLMM results in accounting for session-level differences. The predictability biomarker is computed based on multi-modal physiological and behavioral data. Each arrow originates from the independent variable and points towards the dependent variable. The blue arrow indicates negative correlations, while the red arrow indicates positive correlations. Dashed lines indicate insignificant correlations. · $P < 0.1, *P < 0.05, **P < 0.001$, n = 10.

2.6 Team Performance and Subjective Ratings of Team Members

Familiarity among team members has been demonstrated to be positively correlated with team performance (17; 13). Our experiment focused on a collaborative distributed control task observed a similar pattern (Fig. 4). Specifically, familiarity among co-pilots is positively correlated with team performance, indicating that as familiarity increases, teams perform better on both sub-tasks and in achieving their long-term goals ($\beta = 0.18, P = 0.074$). This finding emphasizes the importance of building familiarity within teams, as it appears to enhance their ability to work cohesively and effectively toward shared goals.

Interestingly, the helpfulness rating of team members is significantly negatively correlated with team performance ($\beta = -0.32, P < 0.001$). This suggests that higher helpfulness ratings may reflect a greater reliance on teammates for support, which could reduce individual autonomy or efficiency, potentially detracting from the overall team performance. Conversely, lower helpfulness ratings may indicate a more balanced contribution from all members, optimizing team efficiency toward achieving shared objectives. Together, the subjective ratings of familiarity and helpfulness reveal a nuanced relationship between team dynamics and performance. While familiarity fosters cohesion and shared understanding, perceptions of helpfulness may introduce dynamics that negatively impact team performance. These insights highlight the complex interplay between subjective perceptions and performance, offering valuable guidance for designing and optimizing collaborative teams in distributed control tasks.

3 Discussion

In this study, we conducted a team-based collaborative virtual reality (VR) experiment and demonstrated a novel multi-modal biomarker that directly correlates with team performance. Specifically, we demonstrated that a biomarker measuring the predictability of teammate behavior is better correlated with team performance. This

biomarker is derived from integrating multi-modal physiology and behavior of teammates to predict the future behavior of the remaining (i.e. left out) team member. Our predictability biomarker challenges the conventional wisdom that physiological and behavioral synchrony is a robust marker of a high-performing team (18; 2; 45; 13; 23).

Simultaneously collecting and analyzing multi-modal data is crucial for understanding team performance and dynamics. In contrast to executing simple tasks individually, collaborative tasks involve complex dynamics and interactions among team members. Various data modalities, including pupillometry, EEG, speech, and other physiological or behavioral data, have been analyzed individually but not in combination (40; 25; 7; 10). We have developed a cross-modal multi-head attention predictive model that is capable of simultaneously analyzing multi-modal data from multiple team members (Fig. 3 b). This model integrates inputs from multiple data modalities, enabling not only the prediction of future actions of individuals but also the identification of a biomarker that is inversely related to overall team performance (22). This result further demonstrates that different physiological and behavioral measures provide unique information that needs to be integrated to construct biomarkers that better relate to performance.

Our results revealed a positive correlation between our predictability biomarker and team performance. Aligning with the common belief that high-performance teams benefit from predictable actions among members (3), our findings suggest that this is expresses in teammate physiology in a way that leads to enhanced coordination and alignment for achieving higher performance (Fig. 4). While the synchrony of individual modalities among co-pilots showed marginal or insignificant correlation with team performance, combining multi-modal data as input to the predictive model revealed that the predictability of team members' future actions is a stable and reliable biomarker of team performance. This highlights the potential of leveraging predictability as a key metric for understanding and improving team dynamics.

We have focused primarily on using predictability as a key indicator of team performance in collaborative tasks involving multiple humans. A pivotal question arising from our research is how we may practically leverage the predictive abilities of team members to enhance team dynamics and performance. This capability can facilitate collaboration between humans or, potentially, teaming between humans and artificial intelligence (AI) agents (5; 26; 16). Our findings lay the groundwork for innovative teaming strategies, fostering enriched and more productive collaborations.

4 Methods

4.1 Participants

Fifty-four healthy human participants (age = 23.67 ± 3.34 year (mean $\pm$ standard deviation); 27 females, 27 males) voluntarily participated in the three experiments. These participants were divided into 18 triad teams, and each team participated in three sessions on different days. Due to incomplete sessions, data from one team were omitted from the final data set. Data from four teams were omitted from the pupil size analysis due to invalid pupil size recordings of one or more co-pilots. EEG data

from nine teams were excluded from the analysis due to error-prone recordings identified during preprocessing. Similarly, speech event data from ten teams were excluded because the speech event detection algorithm failed to extract speech events from one or more participants within the team. No participants or teams dropped out of the experiment due to motion sickness or other symptoms related to virtual reality. All participants had normal or corrected-to-normal visual acuity and gave informed consent before participating in each experiment. Human subject protocols were approved by the Columbia University Institutional Review Board.

4.2 Virtual Environment

The virtual environment was built using *Unreal Engine 4.25.4*. The four main reactive objects in the virtual environment were 1. a spacecraft, 2. a countdown timer, 3. the rings, and 4. the Earth. As shown in Fig.1 b, three viewing windows with different shapes and at different positions were placed at the front of the spacecraft. Each subject in the triad team was assigned to look through one window, and the degree of freedom the subject controls was fixed per experiment session, corresponding to its respective window. The ThrustPilot, who controlled the speed of the spacecraft, had the largest unobstructed field of view, which was located at the bottom of the spacecraft. The YawPilot, who controlled the left-right spacecraft movement, was located at the top-left of the spacecraft. The PitchPilot had a viewing window on the top-right and controlled the up-down movement of the spacecraft. Because the positions and shapes of the windows were different, subjects with different roles had partial and biased views of the environment. The field of view of the virtual camera of each co-pilot is 80°in Unreal Engine.

A countdown timer bar was displayed at the bottom of each window to indicate the remaining time for each trial. Initially, the timer bar was completely black. As time elapsed, the black portion of the bar gradually decreased, revealing an increasing white segment. This white segment represented the time that had passed and was inversely proportionate to the black portion, which showed the remaining time. Each trial had a maximum duration of 55 seconds. The timer would automatically stop and reset if the team either successfully navigated through all the rings and approached Earth, or failed to pass through any ring during the trial. Despite this, with the default speed of the spacecraft, teams would require at least 60 seconds to pass through all rings in a trial, presenting a significant challenge and requirement for active participation and collaboration with the ThrustPilot.

The rings were transparent red toruses that represented a trial's reentry path. At the beginning of each trial, a sequence of fifteen rings was generated, spaced equally but positioned at varying horizontal and vertical coordinates. The distance between any two adjacent rings was $50,000$ units in *Unreal Engine*. The Earth was positioned at the end of the path with $50,000$ units from the final ring. The trial ended when the spacecraft, operated by the team of participants, successfully navigated through all rings and stopped in front of Earth. Upon successful completion, the term "Successful" will be displayed on each participant's head-mounted display (HMD) for one second. Subsequently, a new trial will automatically be started.

4.3 Apparatus

In all experiments, each participant was equipped with VIVE Pro Eye head-mounted displays (HTC Corporation; resolution: 1440×1600 pixels per eye; refresh rate: 90 Hz), and an EEG device with 20 electrodes was placed in accordance with the international 10-20 system (Advanced Brain Monitoring B-Alert X24; sample rate: 256 Hz). A USB microphone was set in front of each subject to enable communication between subjects, and Mumble (version 1.4.230) was running locally on each desktop. We used LabRecorder (version 1.14.0) to collect the multi-modal data. Each head-mounted display is connected to a desktop with an Intel Core i9 CPU and an NVIDIA RTX 2070 Super GPU. The three desktops were connected to a local, secure WiFi network with a 2.6 Gbps router using client-server network protocols to communicate. The server was another desktop with an Intel Core i9 CPU and an NVIDIA RTX 2080 Super GPU.

4.4 Procedure

In each experiment, three participants arrived at the lab and watched an instructional video before the first session. Following the setup of the EEG devices, participants were escorted to three separate EEG recording chambers designed to block sound and electrical noise. These chambers were additionally acoustically shielded with 2-inch thick soundproofing foam to prevent echoes and minimize noise interference. We assisted the participants in setting up head-mounted displays and remote controllers.

Individual eye calibration commenced once each participant was fully equipped and settled. The calibration was conducted using the VIVE Pro Eye system. Each experiment began with five pilot trials following eye calibration, allowing subjects to familiarize themselves with the task environment before the commencement of data collection. A trial was terminated when the team failed to pass a ring due to a crash or a miss or if the time limit was exceeded. After the pilot trials, participants were notified via headphones that the experiment had officially started.

Each team participated in three repeated sessions of the same experiment. Each session was spaced at least 24 hours apart, and no participant had participated in nor had familiarity with the task before their first session. Within each experiment session, roles were randomly assigned to the subjects. After each experimental session, all participants were asked to complete a post-task questionnaire separately (see the post-task survey in Post Task Survey for details).

4.5 Data Preprocessing

We implemented different pre-processing methods for various data modalities. For pupil size data, we first detected and removed blinks and artifacts. Then, we applied linear interpolation and Z-scored the pupil size data. This was followed by averaging the pupil size between the left and right eyes and a fourth-order Butterworth lowpass filter to remove high-frequency noise.

The EEG pre-processing included filtering the raw EEG data using fourth-order Butterworth bandpass filters with bands 0.5 Hz-100 Hz (MNE 1.6.1 (14)). Manual

bad channel rejection was conducted to remove error-prone channels in each recording. Then we performed Independent Component Analysis (ICA) (24) and used the Multiple Artifact Rejection Algorithm (MARA) (47) to separate and reject artifact components.

Remote controller actions were first down-sampled to 60 Hz. Next, values greater than 0.5 were assigned a value of 1, values less than -0.5 were assigned a value of -1, and values between -0.5 and 0.5 were assigned a value of 0.

The speech preprocessing involved three steps. First, we applied the noise reduction function (35) to the speech recordings from each subject to remove background noise. Next, we used a simple voice activity detection function to extract speech events. Finally, the speech events were down-sampled to 60 Hz. All data modalities were then epoched based on the relative time to the respective rings and saved for analysis.

4.6 Post Task Survey

After each experimental session, all participants were asked to complete a survey comprised of demographic and subjective questions. In this study, our analysis concentrated on two specific subjective questions:

1. How helpful was each of your teammates in reaching the final goal?
2. How well did you know each of your team members before today?

Each participant was required to select one of three possible answers for each question that concerned every other team member, excluding themselves. These answers were scaled as 0 = Not at all, 1 = A little, and 2 = Very well. The responses to the helpfulness (Question 1) and familiarity (Question 2) questions were assessed based on the team and the specific experiment session. The helpfulness and familiarity scores ranged from 0 to 12 for each team. A score of 0 indicated that all three participants rated 'Not at all' for each of the other two team members. In contrast, a score of 12 indicated that every participant rated 'Very well' for their teammates.

4.7 Pupil Size, Remote Controller Action, and Speech Event Synchronies

This study computed the inter-subject correlation (ISC) across the three subjects using their pupil sizes, remote controller actions, and speech events. For each experiment session, we computed the Pearson Correlation Coefficient (r) between each pair of participants, participant a and participant b, with their distinct roles within the same team, for one data modality at a time, using (1)

$$r_{a,b} = \frac{\sum_{i=1}^{n}(a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \overline{a})^2 \sum_{i=1}^{n}(b_i - \overline{b})^2}}. \quad (1)$$

n was the length of one epoch of data. The team ISC in one epoch r_{epoch} was the averaged ISC across three co-pilots.

4.8 EEG ISC

To assess inter-brain synchronization, we computed ISC using Correlation Component Analysis (CorrCA) (29). This approach involved utilizing linear combinations of EEG channels or EEG signals with other data modalities as separate channels that maximize the ISC on the data obtained from subjects within the same team. In our study, we employed an improved version of CorrCA to compute the correlation between multiple subjects within the same team while performing a collaborative control task. The EEG signals of each subject contained 20 channels, and the approach finds a weight vector w that maximizes the Pearson Correlation between subjects in the team.

The weight vector w determines which linear combination of different channels provided the most significant correlation among team members. Given the EEG signals of the three subjects, denoted as X_1, X_2, and X_3, where $X_n \in \mathbb{R}^{D\times T}$ with D representing the number of channels and T representing the number of time steps in an epoch, the weight vector w could be computed by:

$$\begin{aligned} &w = argmax_w(\frac{w^T R_{12} w}{\sqrt{w^T R_{11} w}\sqrt{w^T R_{22} w}}); \\ &\text{where } R_{ij} = \frac{1}{T} X_i X_j^T \end{aligned} \tag{2}$$

We defined the within subject covariance as $R_w = \sum_i^N R_{ii}$ and between subject covariance as $R_b = \sum_i^N \sum_{j,j>i}^N R_{ij}$. Here, $N = 3$ denoted the number of subjects in each experiment. We computed the eigenvectors e_k of $R_w^{-1} R_b$ and ranked the eigenvectors in descending order based on the corresponding eigenvalues. Hence, the ISC was the maximum value of the strengths of correlations C_k, where

$$C_k = \frac{e_k^T R_b e_k}{e_k^T R_w e_k}. \tag{3}$$

4.9 The Generative Forecasting Model

The predictive model we implemented was a multi-head attention-based neural network that tracked relationships between events in data within the time domain. Fig. 3 b illustrates the structure of the model. The input to the model included the team's spacecraft trajectory along with the behavioral and physiological data of two participants. The transformer model utilized both encoders and decoders discussed in the original transformer model (44). The 8-head attention layers in the encoder and the masked 8-head attention layers in the decoder were implemented as follows:

$$\begin{aligned} &MultiHead(Q, K, V) = Concat(head_1, ..., head_4)W^O, \\ &\text{where } head_n = Attention(QW_n^Q, KW_n^K, VW_n^V), \\ &[W_n^Q, W_n^K] \in \mathbb{R}^{d_m \times d_k}, W_i^V \in \mathbb{R}^{d_m \times d_v}, W^O \in \mathbb{R}^{hd_v \times d_m} \end{aligned} \tag{4}$$

We used $d_k = d_v = d_m/h = 64$ in this work. The *Attention* function took a set of queries as a matrix Q, the keys matrix K, and the values matrix V. The output of the *Attention* layer was:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{5}$$

All training and testing were conducted on a single NVIDIA RTX A6000 GPU, utilizing CUDA version V12.2.140. To further validate our model, we monitored metrics such as loss and accuracy during the training phase and utilized a validation dataset to assess performance periodically.

4.10 Model Evaluation

We evaluated the predictive model's performance by computing the Pearson correlation coefficient r between the prediction and the target. To do so, we first computed the correlation of each individual using (1), where a was the concatenated target actions, and b was the concatenated model predictions.

4.11 Predictability as a Biomarker

The predictive model we developed generates predicted future actions for each co-pilot based on the behavioral and physiological data of the other two co-pilots. These predictions are then correlated with the co-pilots' actual actions to compute a unique correlation score for each individual. We employ (1) to calculate the holistic team biomarker, which averages the predictability scores across the three co-pilots.

4.12 Generalized Linear Mixed-effect Model

As an extension to the generalized linear model (GLM), the linear predictors of the generalized linear mixed-effects model (GLMM) contained random effects in addition to the usual fixed effects (4). We used the GLMM in Python (statsmodels (38)) to investigate the relationship between varied variables with team difference considered as random-effect (37). The final regression formula of each model was listed in supplementary materials in general form:

$$y = X\beta + Z\mu + \varepsilon, \tag{6}$$

where y is the outcome variable. X represents the predictor variables. β is a column vector of the fixed-effects regression coefficients, and Z is the design matrix for the random effects (the random complement to the fixed X). μ is a vector of the random effects (the random complement to the fixed β), and ε is a column vector of the residuals. The supplementary materials list all the details, including all models' fixed and random effects.

5 Acknowledgements

This work was supported by funding from the Army Research Laboratory's STRONG Program (W911NF-19-2-0139, W911NF-19-2-0135, W911NF-21-2-0125) the National Science Foundation (IIS-1816363, OIA-1934968) the Air Force Office of Scientific Research (FA9550-22-1-0337) and a Vannevar Bush Faculty Fellowship from the US Department of Defense (N00014-20-1-2027).

References

[1] (1995) Apollo 13. Starring Tom Hanks, Kevin Bacon, and Bill Paxton

[2] Abney DH, Paxton A, Dale R, et al (2015) Movement dynamics reflect a functional role for weak coupling and role structure in dyadic problem solving. Cognitive processing 16:325–332

[3] Bradley BH, Baur JE, Banford CG, et al (2013) Team players and collective performance: How agreeableness affects team performance over time. Small Group Research 44(6):680–711

[4] Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. Journal of the American statistical Association 88(421):9–25

[5] Chen B, Vondrick C, Lipson H (2021) Visual behavior modelling for robotic theory of mind. Scientific Reports 11(1):424

[6] Dabbish L, Kraut R, Patton J (2012) Communication and commitment in an online game team. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 879–888

[7] Dias RD, Zenati MA, Stevens R, et al (2019) Physiological synchronization and entropy as measures of team cognitive load. Journal of biomedical informatics 96:103250

[8] Dikker S, Wan L, Davidesco I, et al (2017) Brain-to-brain synchrony tracks real-world dynamic group interactions in the classroom. Current biology 27(9):1375–1380

[9] Dmochowski JP, Sajda P, Dias J, et al (2012) Correlated components of ongoing eeg point to emotionally laden attention–a possible marker of engagement? Frontiers in human neuroscience 6:112

[10] Dunbar TA, Gorman JC (2020) Using communication to modulate neural synchronization in teams. Frontiers in Human Neuroscience 14:332

[11] Edmondson A (1999) Psychological safety and learning behavior in work teams. Administrative science quarterly 44(2):350–383

[12] Faller J, Cummings J, Saproo S, et al (2019) Regulation of arousal via online neurofeedback improves human performance in a demanding sensory-motor task. Proceedings of the National Academy of Sciences of the United States of America 116(13):6482 – 6490. https://doi.org/10.1073/pnas.1817207116

[13] Gordon I, Gilboa A, Cohen S, et al (2020) Physiological and behavioral synchrony predict group cohesion and performance. Scientific reports 10(1):8484

[14] Gramfort A, Luessi M, Larson E, et al (2013) MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience 7(267):1–13. https://doi.org/10.3389/fnins.2013.00267

[15] Hansen A, Larsen KB, Nielsen HH, et al (2020) Asymmetrical multiplayer versus single player: Effects on game experience in a virtual reality edutainment game. In: International Conference on Augmented Reality, Virtual Reality and Computer Graphics, Springer, pp 22–33

[16] Harris-Watson AM, Larson LE, Lauharatanahirun N, et al (2023) Social perception in human-ai teams: Warmth and competence predict receptivity to ai teammates. Computers in Human Behavior 145:107765

[17] Harrison DA, Mohammed S, McGrath JE, et al (2003) Time matters in team performance: Effects of member familiarity, entrainment, and task discontinuity on speed and quality. Personnel Psychology 56(3):633–669

[18] Henning RA, Boucsein W, Gil MC (2001) Social–physiological compliance as a determinant of team performance. International Journal of Psychophysiology 40(3):221–232

[19] Jang Y, Ryu S (2011) Exploring game experiences and game leadership in massively multiplayer online role-playing games. British Journal of Educational Technology 42(4):616–623

[20] Kauppi JP, Jääskeläinen IP, Sams M, et al (2010) Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. Frontiers in neuroinformatics 4:669

[21] Keles U, Dubois J, Le KJ, et al (2024) Multimodal single-neuron, intracranial eeg, and fmri brain responses during movie watching in human patients. Scientific Data 11(1):214

[22] Madsen AG, Lehn-Schiøler WT, Jónsdóttir Á, et al (2023) Concept-based explainability for an eeg transformer model. In: 2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, pp 1–6

[23] Madsen J, Parra LC (2022) Cognitive processing of a common stimulus synchronizes brains, hearts, and eyes. PNAS nexus 1(1):pgac020

[24] Makeig S, Bell A, Jung TP, et al (1995) Independent component analysis of electroencephalographic data. Advances in neural information processing systems 8

[25] McCraty R (2017) New frontiers in heart rate variability and social coherence research: techniques, technologies, and implications for improving group dynamics and outcomes. Frontiers in public health 5:267

[26] Metcalfe JS, Perelman BS, Boothe DL, et al (2021) Systemic oversimplification limits the potential for human-ai partnership. IEEE Access 9:70242–70260

[27] Moore SM, Geuss MN (2020) Familiarity with teammate's attitudes improves team performance in virtual reality. PloS one 15(10):e0241011

[28] Morgan PJ, Pittini R, Regehr G, et al (2007) Evaluating teamwork in a simulated obstetric environment. The Journal of the American Society of Anesthesiologists 106(5):907–915

[29] Parra LC, Haufe S, Dmochowski JP (2018) Correlated components analysis-extracting reliable dimensions in multivariate data. arXiv preprint arXiv:180108881

[30] Pearsall MJ, Ellis AP (2006) The effects of critical team member assertiveness on team performance and satisfaction. Journal of Management 32(4):575–594

[31] Pobiedina N, Neidhardt J, Moreno MdCC, et al (2013) On successful team formation: Statistical analysis of a multiplayer online game. In: 2013 IEEE 15th Conference on Business Informatics, IEEE, pp 55–62

[32] Poulsen AT, Kamronn S, Dmochowski J, et al (2017) Eeg in the classroom: Synchronised neural recordings during video presentation. Scientific reports 7(1):43916

[33] Reinero DA, Dikker S, Van Bavel JJ (2021) Inter-brain synchrony in teams predicts collective performance. Social cognitive and affective neuroscience 16(1-2):43–57

[34] Rerup C (2001) "Houston, we have a problem": Anticipation and improvisation as sources of organizational resilience. Snider Entrepreneurial Center, Wharton School Philadelphia, PA

[35] Sainburg T, Thielk M, Gentner TQ (2020) Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. PLoS computational biology 16(10):e1008228

[36] Sapienza A, Zeng Y, Bessi A, et al (2018) Individual performance in team-based online games. Royal Society open science 5(6):180329

[37] Seabold S, Perktold J (2010) statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference

[38] Skipper S, Josef P (2010) statsmodels: Econometric and statistical modeling with python. 9th Python in Science Conference

[39] Špiláková B, Shaw DJ, Czekóová K, et al (2020) Getting into sync: Data-driven analyses reveal patterns of neural coupling that distinguish among different social exchanges. Human brain mapping 41(4):1072–1083

[40] Stevens R, Galloway T, Gorman J, et al (2016) Toward objective measures of team dynamics during healthcare simulation training. Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care 5(1):50–54. https://doi.org/10.1177/2327857916051010, URL https://doi.org/10.1177/2327857916051010, https://doi.org/10.1177/2327857916051010

[41] Szymanski C, Pesquita A, Brennan AA, et al (2017) Teams on the same wavelength perform better: Inter-brain phase synchronization constitutes a neural substrate for social facilitation. Neuroimage 152:425–436

[42] Van der Vaart T, Van Donk DP (2008) A critical review of survey-based research in supply chain integration. International journal of production economics 111(1):42–55

[43] Varlet M, Filippeschi A, Ben-Sadoun G, et al (2013) Virtual reality as a tool to learn interpersonal coordination: Example of team rowing. Presence 22(3):202–215

[44] Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Advances in neural information processing systems 30

[45] Vicaria IM, Dickens L (2016) Meta-analyses of the intra-and interpersonal outcomes of interpersonal coordination. Journal of Nonverbal Behavior 40:335–361

[46] Weissker T, Froehlich B (2021) Group navigation for guided tours in distributed virtual environments. IEEE Transactions on Visualization and Computer Graphics 27(5):2524–2534

[47] Winkler I, Haufe S, Tangermann M (2011) Automatic classification of artifactual ica-components for artifact removal in eeg signals. Behavioral and brain functions 7:1–15

[48] Wohltjen S, Toth B, Boncz A, et al (2023) Synchrony to a beat predicts synchrony with other minds. Scientific Reports 13(1):3591

[49] Xie H, Karipidis II, Howell A, et al (2020) Finding the neural correlates of collaboration using a three-person fmri hyperscanning paradigm. Proceedings of the National Academy of Sciences 117(37):23066–23072