ARI-SS-SR 2024-03
Scientist Series
Scientific Report

February 2024

Cohesion and Organization through Networked Computation in Overly Risky and Distinctly Isolated Areas (CONCORDIA)

Jeremy Gottlieb
Noam Benkler
Joan Zheng
Smart Information Flow Technologies, LLC

Pete Roma
Naval Health Research Center

Nathanael L. Keiser
U.S. Army Research Institute

United States Army Research Institute for the Behavioral and Social Sciences

United States Army Research Institute
for the Behavioral and Social Sciences

A Directorate of the Department of the Army
Deputy Chief of Staff, G1

Authorized and approved:

SCOTT B. SHADRICK, Ph.D.
Acting Director

Research accomplished under contract
for the Department of the Army by:

Smart Information Flow Technologies, LLC

Research Unit Chief
Dr. Brian Crabb (Fort Cavazos Research Unit)

Team Leader
Dr. Christopher Vowels

Technical Reviewers
Nicole T. Harrington, U.S. Army Research Institute
Kathleen Darbor, U.S. Army Research Institute

DISPOSITION

This Technical Report has been submitted to
the Defense Technical Information Center (DTIC).

REPORT DOCUMENTATION PAGE

1. REPORT DATE (Month Year)	2. REPORT TYPE	3. DATES COVERED (Month Year)	
		START DATE	END DATE
February 2024	Final	September 2021	February 2023

4. TITLE AND SUBTITLE

Cohesion and Organization through Networked Computation in Overly Risky and Distinctly Isolated Areas (CONCORDIA)

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. COOPERATIVE AGREEMENT NUMBER
W911NF-21-C-0052		

5d. PROGRAM ELEMENT NUMBER	5e. PROJECT NUMBER	5f. TASK NUMBER	5g. WORK UNIT NUMBER
0602785A	790	865	

6. AUTHOR(S)

Gottlieb, Jeremy, Benkler, Noam, Zheng, Joan., Roma, Pete, and Keiser, Nathanael L.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Smart Information Flow Technologies
319 1st Ave North, Suite 400
Minneapolis, MN 55401-1689

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
U.S. Army Research Institute for the Behavioral and Social Sciences 6000 6th Street (Bldg. 1464 / Mail Stop: 5610) Fort Belvoir, Virginia 22060	ARI	Scientific Report 2024-03

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution Statement A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

ARI POC: Dr. Nathanael Keiser, Fort Cavazos Research Unit

14. ABSTRACT

Cohesion and Organization through Networked Computation in Overly Risky and Distinctly Isolated Areas (CONCORDIA) was an attempt to develop a set of algorithms for assessing team cohesion via wearable devices collecting physiological data from team members while they performed tasks designed specifically to assess team cohesion – COHESION and CubeCrusher. The algorithms utilized a variety of machine learning techniques to identify when a team member's behavior started to deviate from that of the rest of their team, as well as to predict performance metrics. The results, while not conclusive, indicate that there is a significant probability that algorithms with a high level of accuracy and robustness can be developed. We also developed, in concurrence with the machine learning algorithms, a prototype interface for displaying team cohesion in real time according to physiological data. We provide suggestions for future research directions that could make such a system effective.

15. SUBJECT TERMS

Team Cohesion, Wearable Devices, Extreme Environments

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
Unclassified	Unclassified	Unclassified	Unlimited Unclassified	44

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER (Include area code)
Brian T. Crabb	254-288-3833

COHESION AND ORGANIZATION THROUGH NETWORKED COMPUTATION IN OVERLY RISKY AND DISTINCTLY ISOLATED AREAS (CONCORDIA)

EXECUTIVE SUMMARY

Research Requirement:

Army leaders and Soldiers need quantifiable metrics and measures for team cohesion that: (1) are explainable, (2) allow users to drill-down into performance and success factors, and therefore, (3) are beyond the optimization scores typically generated by existing automated systems. One potential avenue for increasing the Army's ability to maximize team performance is to develop methods for autonomously assessing team cohesion in near-real-time using data streams generated by wearable biometric measuring devices. The purpose of Cohesion and Organization through Networked Computation in Overly Risky and Distinctly Isolated Areas (CONCORDIA) was to assess the scientific feasibility of such a system by designing prototype models for assessing team cohesion using data collected from teams working in isolation at the HI-SEAS extreme environment habitat.

Procedure:

The CONCORDIA utilized data collected from already validated team cohesion assessment tasks, as well as physiological data collected from Empatica E4 devices worn by the team members performing the COHESION and CubeCrusher tasks. These data were used to develop machine learning models of team cohesion within those tasks based on the physiological data from the Empatica devices. Once initial models were constructed from already collected data, we gathered further data from Soldiers at an Army Combat LifeSaver course wearing the Empatica devices and evaluated it using the same team cohesion models. These efforts were coupled with the development of a prototype user-friendly interface for displaying team cohesion metrics in real time.

Findings:

While not completely conclusive, our results indicate that it is possible to generate an assessment of team cohesion from physiological data gathered by wearable devices. Further research is needed to develop better longitudinal metrics of team cohesion, to understand the role of context, and to test developed algorithms in settings with higher fidelity, but while maintaining enough control to reach valid conclusions. Our user interface leveraged a design based on stop-light, radial diagrams; however, this solution is a prototype because it has not been tested in an operational context nor subjected to a user evaluation study.

Utilization and Dissemination of Findings:

The findings presented in this report primarily serve as a roadmap for developing more accurate and robust system to assess team cohesion from wearable biometrics. As such, this report is directly applicable to researchers who study team cohesion in the Army. The results and

prototype user interface are also generalizable to Army leaders interested in better tracking team cohesion in near-real-time in the field.

COHESION AND ORGANIZATION THROUGH NETWORKED COMPUTATION IN OVERLY RISKY AND DISTINCTLY ISOLATED AREAS (CONCORDIA)

CONTENTS

Page

APPENDICES

LIST OF TABLES

LIST OF FIGURES

Page

COHESION AND ORGANIZATION THROUGH NETWORKED COMPUTATION IN OVERLY RISKY AND DISTINCTLY ISOLATED AREAS (CONCORDIA)

Introduction

The goal of Cohesion and Organization through Networked Computation in Overly Risky and Distinctly Isolated Areas (CONCORDIA) is to establish the scientific feasibility of developing metrics for team cohesion based solely on data gathered remotely from wearable devices. We attempted to achieve this goal by leveraging data collected via our MEDULLA (Monitoring Emergent Duress Under Long-Lasting Allostasis) program at the HI-SEAS (Hawai‘i Space Exploration Analog and Simulation) space analog site (www.hi-seas.org). This dataset contains validated measures of team cohesion, such as NASA’s COHESION game and our CubeCrusher game, as well as physiometric data, surveys, and linguistic data from both game play and participant journals. The dataset also contains data from the Empatica E4 wristband. This allowed us to align the data from the wearable devices with the team cohesion signals in the validated measures to identify what metrics from the wearables, if any, can be used to accurately assess team cohesion.

Team Cohesion

Teams constitute two or more people who work interdependently to achieve a common set of goals (Baker & Salas, 1997). Given the continued reliance on teams in organizations, there has been a spate of research aimed at identifying and assessing factors that contribute to team performance. The focus of CONCORDIA is on team processes, which refer to cognitive, verbal, and behavioral activities among team members aimed at achieving collective goals (Marks et al., 2001). We focus on one particular team process variable that has received extensive empirical assessment: cohesion. Team cohesion is typically conceptualized as a multidimensional construct, constituting task and social elements, along with group pride (Beal et el., 2003).

We focus on team task cohesion because it is theoretically most proximal to team performance (Grossman et al., 2022). Team task cohesion reflects the degree of shared commitment in accomplishing team-based tasks, whereas social cohesion and group pride encompass the attachment among team members and the importance they place on being a part of the team (Beal et al., 2003). There is consistent meta-analytic evidence that team task cohesion is positively associated with team performance (Beal et al., 2003; Castaño et al., 2013; Chiocchio & Essiembre, 2009; Evans & Dion, 1991; Grossman et al., 2022; Gully et al., 1995).

Sensing Technology to Measure Team Task Cohesion

Team task cohesion is most commonly assessed using Likert-based measures, administered to individual team members, and then aggregated to the team level. However, there is a burgeoning literature on the use of sensing technology to measure team constructs (e.g., Chaffin et al., 2017; Kozlowski et al., 2016; Luciano et al., 2018; Matusik et al., 2019). Sensing technology includes external sensors (e.g., video cameras, depth cameras) and body-worn sensors (e.g., vital sensors, sociometric badges, smartphones) capable of collecting information about: (a) movement, position, gestures, and facial expressions, (b) speech patterns and content,

and (c) physiological responses (Langer et al., 2019; Luciano et al., 2018). The literature to date clearly indicates that sensing technology can be used to measure various psychological constructs of interest (Luciano et al., 2018).

The focus of CONCORDIA is on the use of a body-worn sensor to collect physiological response data to measure team task cohesion. To be clear, this is not the first attempt at using body-worn sensors to assess team constructs (also termed *team physiological dynamics*) nor is it the first attempt to measure cohesion based on the same (Kazi et al., 2021). Consistent with much of the literature to date, we expected that teams that are higher on cohesion would display physiological synchrony, meaning that their physiological responses are increasingly similar over time (Guastello, 2016, 2017; Strang et al., 2014).

There are, however, limitations of the existing literature, which serves as the impetus for CONCORDIA. First, only a few researchers have studied the use of physiological metrics to assess team cohesion (Mønster et al., 2016; Strang et al., 2014). Coupled with limited empirical study is mixed support for the expected positive relationship between physiological synchrony (i.e., cardiac interbeat interval, postural sway, skin conductance, and smiling) and self-reported cohesion. Second, many of the existing sensors record data continuously, but these data are stored on the device to be later downloaded and then analyzed (Kozlowski et al., 2016). That type of system is amenable for research purposes but cannot be used to provide the type of real-time information that is useful in an operational context. Third, the application of body-worn sensors is typically relegated to controlled, laboratory-type environments (e.g., Strang et al., 2014). In comparison, military operations are an extremely hectic, loud, and hostile environment (i.e., extreme environment; Driskell et al., 2018). Thus, there remain unknowns about the feasibility of relying on existing sensors in a real-world military context, especially the types of extreme environments expected in future operations.

Research Goals

At a high level, CONCORDIA focused on answering three big questions:

1. Can team task cohesion (hereafter referred to as “team cohesion”) be assessed solely from data generated by wearable devices?

To answer this question, we collected a wide variety of data from the Empatica E4 wristband, including heart rate, respiration, skin temperature, and electrodermal activity. Along with the wearable data, we collected a variety of validated team cohesion metrics. In particular, we collected data from sub-teams and from the full MEDULLA crew on the COHESION team exercise developed by our consultant, Dr. Pete Roma, while he was at NASA. We also collected data on a team game that Smart Information Flow Technologies, LLC (SIFT) developed called CubeCrusher, a virtual reality, Tetris-like game. COHESION and CubeCrusher, combined with the data from the wearable devices, gives us a continuous, real-time stream of data about team cohesion and performance while team members are engaged in a task. Again, consistent with the extant literature, we expect that physiological synchrony among team members is associated with higher team cohesion and performance.

2. Are there signals from the wearables that predict the trajectory of team cohesion?

Being able to assess team cohesion in near-real-time is extremely useful. Being able to analyze the data coming from wearable devices, collate together the multiple streams of data from a single team, and predict the likelihood that team cohesion, and thus performance, may break down in the future is even better. This ability would allow Army leaders to intervene with a poorly performing team before they negatively impact mission performance.

3. Can these techniques be applied to teams operating in extreme environments?

Data coming from operations in extreme environments will be noisy and incomplete. This is due both to communications interference with the signals from the wearable devices to the system processing the data streams, as well as the noise that comes from using the devices in these kinds of extreme situations.

Thus far, we have found that the answer to all three of these questions is "Possibly." As will be described in later sections, we had moderate success predicting behaviors that we hypothesized as being related to team cohesion, as well as predicting performance metrics that are believed to be indicative of team cohesion. However, in the process of building our models and generating results, we identified a number of contextual confounds that indicate it may be prohibitively difficult to construct a one-size-fits-all solution.

Expected Impact

Effectively responding to the challenges of the future battlefield and multi-domain operations will necessitate an understanding of team performance and cohesion in near-real-time. CONCORDIA established some scientific feasibility for evaluating team cohesion solely from wearable biometrics, but more work is needed to develop a deployable solution. CONCORDIA is designed for teams under high-stress, and our target users are in the Army and other military operations, but CONCORDIA could presumably be expanded into telehealth and disease prevention, assessments in organizations, or people analytics.

Team assessments often necessitate time with no perceived benefit to the Soldier, and rely on intrusive and subjective metrics, which often suffer from social desirability biases. CONCORDIA is most impactful to Army leaders through capabilities that allow for: (1) monitoring team functioning via diagnostics, (2) improving team composition via psychometric analytics of individuals, (3) providing insights into root causes of team conflict through data aggregation from larger populations, (4) providing an opportunity to recommend countermeasures before performance decline, and most importantly (5) helping improve team functioning and subsequently optimizing team performance. Broadly, CONCORDIA is impactful in the Army because it provides the basis for assessing more teams, more often, and adopting a continuous improvement process for team interactions.

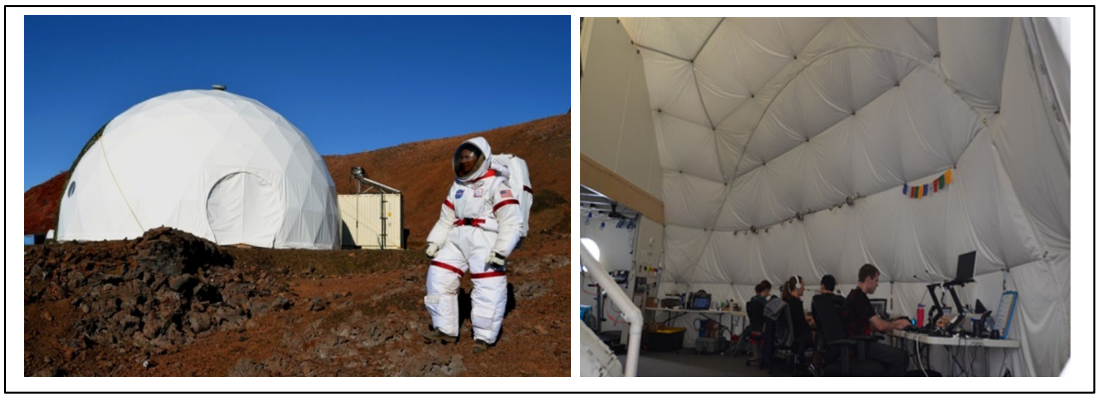
Data Collection — MEDULLA and HI-SEAS

HI-SEAS Habitat

HI-SEAS is a habitat on an isolated, Moon and Mars-like site on the Mauna Loa side of the saddle area on the Big Island of Hawaii at approximately 8200 feet above sea level (Figure 1). HI-SEAS is unique, in addition to its setting in a distinctive analog environment, as: (a) we select the crew to meet our research needs (in serendipitous analogs, such as Antarctic stations, crew selection criteria are not controlled by researchers); (b) the conditions (habitat, mission, communications, etc.) are explicitly designed to be similar to those of a planetary exploration mission; (c) the site is accessible year-round, allowing longer-duration isolated and confined environment studies than at other locations; (d) the Moon and Mars-like environment offers the potential for high-fidelity analog tasks, such as geological field work carried out by human explorers and/or robots.

Figure 1

HI-SEAS Habitat



Living Conditions

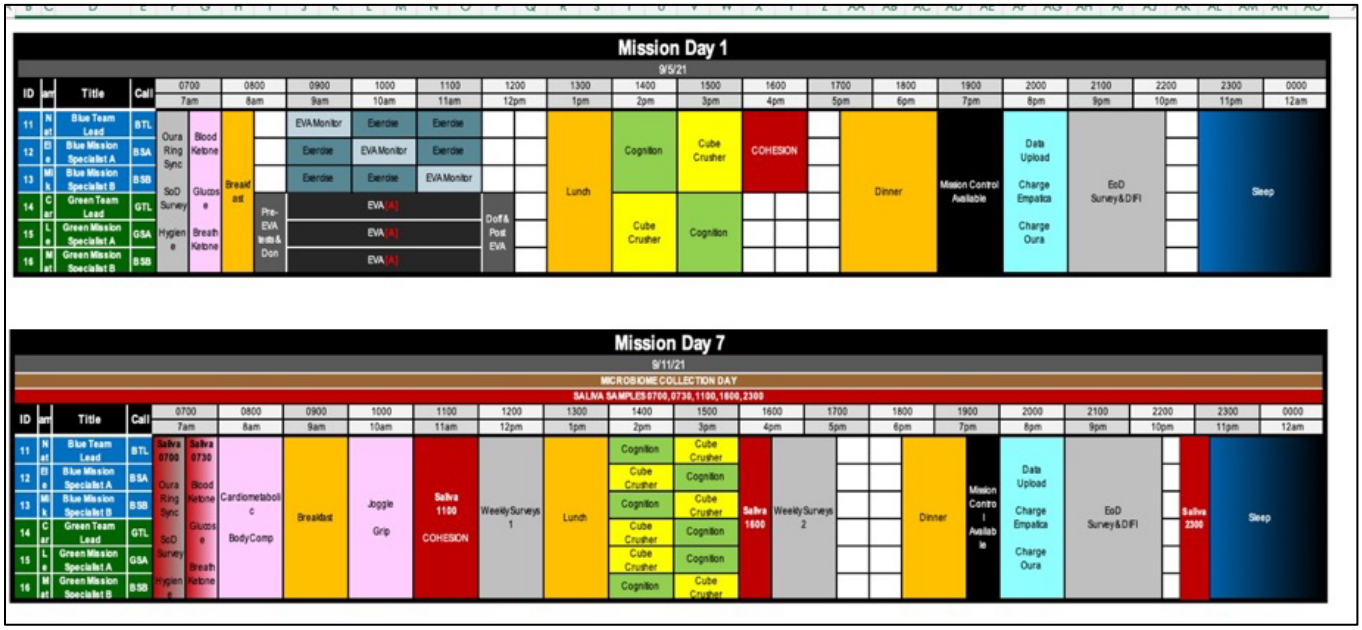
Subjects lived in the habitat continuously for the 28-day duration of a mission. During the course of these four weeks, they engaged in a wide variety of tasks that fall into four broad areas: experimental tasks related to the various experimental sub-protocols being utilized, performance tasks where we can judge team and individual performance under various experimental conditions, habitat maintenance tasks such as cleaning and simple repairs, and life tasks such as eating and sleeping (Figure 2).

Subjects occupied the shared space of the habitat. The total square footage of the living space of the habitat is 1200 square feet. The crew was provided with 400 gallons of water for 1 month, which is water used for everything from drinking, washing, cooking, etc. They were instructed about the proper usage of water to ensure enough was available throughout the mission. The crew's food was also provided at the beginning of the mission and expected to last the entire 28 days. The food was typical of a long-duration, isolated exercise, including dried fruits and vegetables, root vegetables, dried meats, beans, cheeses, and other similar foods.

The habitat was equipped with physical training/fitness equipment such as a treadmill, stationary bicycle, weights, jumping ropes, and six yoga mats, and subjects were instructed to engage in physical exercise according to the mission schedule.

Figure 2

Typical MEDULLA Mission Days



The MEDULLA study design is based on four mixed-gender teams, with three women and three men in each crew. Each crew is further divided into sub-teams (green and blue). Thus, each crew member is part of a 6-person team, as well as a 3-person sub-team. Each sub-team has either a male or female team lead, as well as mission specialist A and mission specialist B. This also means that male and female team leads must co-lead the whole crew as well as manage their two mission specialists. Additionally, we can study if sub-teams naturally take on a collaborative vs. competitive strategy to group living.

Physiological, Behavioral, and Self-Report Measures at MEDULLA

Figure 3 provides an overview of all measures during the four MEDULLA missions, including a timeline for the various pre-, during, and post-mission measurements. By collecting numerous repeated measurements such as the ones mentioned above, we can see the dynamics of the social allostatic load process over time.

Our primary physiological measures are peripherical measures from the EMPATICA E4 wristband. The E4 is a wearable, non-invasive, research device which allows for real-time physiological data collection such as blood volume pulse (BVP) from which heart rate (HR), interbeat intervals (IBIs) in milliseconds, and heart rate variability (HRV) are derived.

Peripheral measures such as HR and HRV are a physiological indicator of adaptive emotion regulation and fluctuations in mental load (Appelhans & Luecken, 2006; Thayer et al., 2009) and can also be used to assess team cohesion (Milosevic et al., 2013). Greater HRV is a physiological indicator of adaptive emotion regulation and a decreased mental load (Nickel &

Nachreiner, 2003). A meta-analysis of HRV and neuroimaging studies by Thayer et al. (2012) links HRV as a marker of stress and health, showing cerebral blood flow increases during emotion control tasks. Thayer et al. found that the automatic nervous system and parasympathetic tone are associated with the regulation of allostatic systems associated with glucose regulation, hypothalamic-pituitary-adrenal (HPA) axis function and inflammatory processes.

Figure 3

Summary Table of Measures Taken During MEDULLA Missions

Measure	Measure Type	Mission Phase					
		Offsite/ Home	Onsite			Offsite/ Home	Modification for M3&M4
		Pre-Training	Pre-Mission	In-Mission	Post-Mission	Follow-Up	
CubeCrusher	Behavioral Task		Daily	Daily	Daily		
COHESION	Behavioral Task		Daily	Daily	Daily		
Cognition	Behavioral Task		Daily	Daily	Daily		
Extravehicular Activity (EVA) Field Exercises	Behavioral Task		Daily	Every 2 Days			
NASA Meaningful Work & Enjoyment Scales (MeWES)	Self-Report	1x	Daily	Daily	Daily		
Neurobehavioral Function	Self-Report		Daily	Daily	Daily		
Sleep	Self-Report		Daily	Daily	Daily		
Profile of Mood States Short Form (POMS-SF)	Self-Report		Daily	Daily	Daily		
MeWES Daily	Self-Report		Daily	Daily	Daily		
Team Performance	Self-Report		Daily	Daily	Daily		
Team Cohesion	Self-Report		Daily	Daily	Daily		
Daily Journal	Self-Report		Daily	Daily	Daily		
HABCOM Team Performance	Self-Report			Daily			
EVA Report	Self-Report			Every 2 Days			
NASA Task Load Index (TLX)	Self-Report			Every 2 Days			
Dynamic Identify Fusion Index (DIFI)	Self-Report	1x	2x	Every 7 Days	1x	1x	1x
Generalized Anxiety Disorder Assessment (GAD-7)	Self-Report	1x	2x	Every 7 Days	1x	1x	1x
Short Stress State Questionnaire (SSSQ)	Self-Report	1x	2x	Every 7 Days	1x	1x	1x
Dyadic Social Interactions Inventory (DSII)	Self-Report		1x	Every 7 Days	1x		
ENRICHD-OP Social Support	Self-Report		1x	Every 7 Days	1x		
Group Living	Self-Report		1x	Every 7 Days	1x		
Psychological Safety	Self-Report		1x	Every 7 Days	1x		
Social Support Questionnaire (SSQ)	Self-Report		1x	Every 7 Days	1x		
Team Processes	Self-Report		1x	Every 7 Days	1x		
Shared Knowledge	Self-Report	1x	2x	Every 14 Days	1x		
Subjective Habitability & Acceptability Questionnaire (SHAQ)	Self-Report	1x	1x	Every 14 Days	1x	1x	1x
Attachment Styles Questionnaire (ASQ)	Self-Report	1x					
Big 5	Self-Report	1x					
Dark Triad	Self-Report	1x					
Demographics	Self-Report	1x					
Deployment Risk and Resilience Inventory 2 (DRRI-2)	Self-Report	1x					
Foursight	Self-Report	1x					
Interpersonal Reactivity Index	Self-Report	1x					
IPIP-NEO-120	Self-Report	1x					
Reading the Mind in the Eye	Self-Report	1x					
Social Desirability Scale (SDS-17)	Self-Report	1x					
Sociosexual Orientation Inventory	Self-Report	1x					
Oura (ring)	Physiology	7 Days	Continuous	Continuous	Continuous	7 Days	
Empatica E4 (watch)	Physiology		Continuous	Continuous	Continuous		
Blood Glucose and Ketone	Physiology	1x	3x	Every 2 Days	3x	1x	n/a for M3/M4
Biosense Breath Ketone	Physiology	1x	3x	Every 2 Days	3x	1x	n/a for M3/M4
Blood Cardio Metabolic Test	Physiology	1x	1x	Every 2 Days	1x	1x	n/a for M3/M4
Body Composition	Physiology		1x	Every 2 Days	1x		n/a for M3/M4
Dexterity	Physiology		1x	Every 2 Days	1x		n/a for M3/M4
Grip Strength	Physiology		1x	Every 2 Days	1x		n/a for M3/M4
Saliva Sampling	Physiology		1 Day	Every 7 Days	1 Day		
Microbiome	Physiology	1x	1x	Every 14 Days	1x	1x	n/a for M4
Taste Perception	Physiology		1x	Every 14 Days	1x		n/a for M3/M4
Visual Acuity	Physiology		1x	Every 14 Days	1x		n/a for M3/M4

HR and IBI are measured by the Empatica E4 device throughout the day, during all tasks, including EVAs, Cognition, COHESION and CubeCrusher. It allows us to do timeseries analysis correlating HR with emotional measures (e.g., language) as well as performance outcomes. The Empatica E4 also measures EDA and skin temperature.

Development of Team Cohesion Metrics

Ground Truth

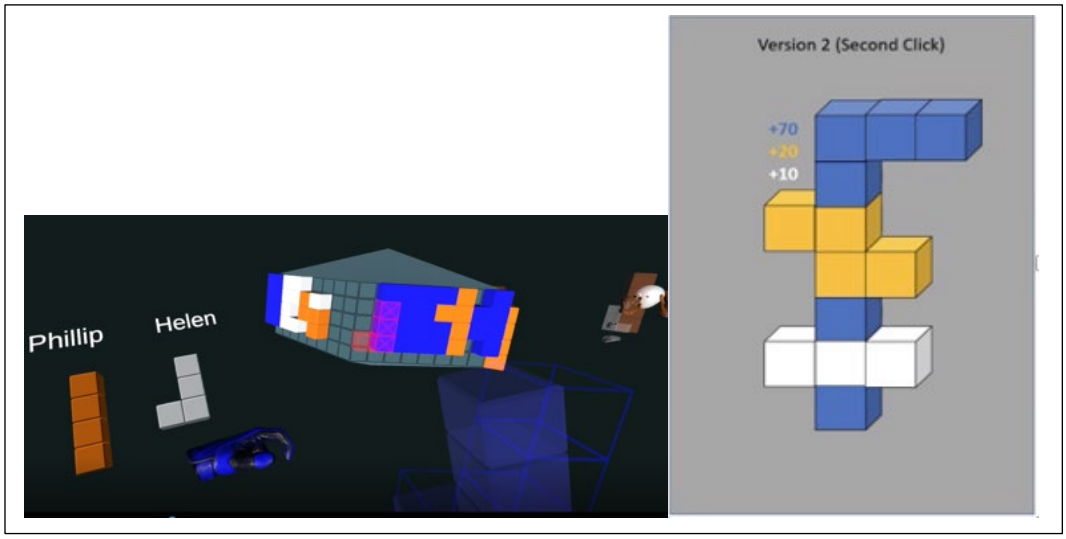
The primary difficulty we encountered early in CONCORDIA was establishing what would serve as the ground truth for evaluating how well we could assess team cohesion solely from physiometric data collected from wearable devices. We collected data from two different tasks designed to assess team cohesion: the validated COHESION game, and the CubeCrusher game developed for the express purpose of assessing cooperation and team dynamics. Both games are described below, but neither served as a perfect ground truth for team cohesion.

CubeCrusher

The CubeCrusher VR game is a Tetris™-like game designed to assess team performance and team dynamics in a controlled setting. To play the game, subjects put on the Vive Cosmos, a virtual reality (VR) headset. Within the VR environment, the game "board" consists of a three-dimensional solid where all the sides are divided into grids (see Figure 4). Tetris-shaped pieces consisting of four cubes appear in the player's hand and they need to place them on the grid, with the goal of having completed columns of filled squares. Every 20 seconds, each player receives a new block that they need place somewhere on the grid. Players can swap pieces with each other and observe any part of the game board, though they can only see the part in front of them at any given time. The game ends when a player cannot place their current piece anywhere on the game board. So, CubeCrusher is similar to Tetris but with some notable alterations to its design and structure (e.g., pieces do not drop) and gameplay (i.e., trades) to incorporate interdependence.

Figure 4

Screenshots of CubeCrusher VR Game and Bonus Points for Cooperation



Crew members played the game in teams of three for 45 minutes at a time, with each subject playing three times per week. The VR headsets include microphones and headphones so that subjects can talk to each other during the game. These conversations were recorded for later linguistic analysis, assessing such things as politeness and emotional responses. We used the following performance measure data:

• Number of columns completed: This is the basic unit of game performance assessment because completing columns is both the goal of the game and the only means by which a game-over state can be avoided.

• Number of players contributing to completed columns: A given player can complete a column all by themselves, or a column can be completed because of players cooperating with each other to correctly place pieces on the board (see Figure 4). The higher the average number of players contributing the completed columns, the more it indicates that players are working together.

• Number of trades: Trading pieces is an indication that players on a team are working together to minimize the likelihood of reaching a game-over state[1]. We are able to link the linguistic analysis from their recorded conversations to specific game actions such as trades to understand the nature of the cooperation that led to a particular trade. For example, the linguistic analysis can identify if players were polite or rude to each other in their pre-trade conversation.

Our expectation was that team cohesion would be indicative of more cooperative behaviors such that either the number of swaps or the number of players contributing to a column would indicate cooperative behaviors, and thus task cohesion at CubeCrusher. To encourage such behaviors, teams and individuals could both earn more points when columns were completed with pieces from multiple players, as in Figure 4.

We identified two confounds in using these metrics. First, participants started to report that using the swapping mechanism was slower than simply asking a teammate to move over to where they were to fill in a gap on the board. While that is certainly cooperative behavior, we have no objective measure of how often that happens. We are working on how to consistently and objectively identify such events from the speech transcripts, since players do not always explicitly request in the same way.

The second is that, through their journal reports, we are aware that some teams experimented with different strategies during different CubeCrusher sessions. For example, teams would discuss before their CubeCrusher session that they would all try to work as fast as possible to finish columns individually, without worrying about swapping or moving around the board to earn bonus points. Thus, while based on our metrics it would appear that the team was not cooperating at all, the a priori agreement on a strategy followed by the in-game commitment to it in fact indicates high task cohesion.

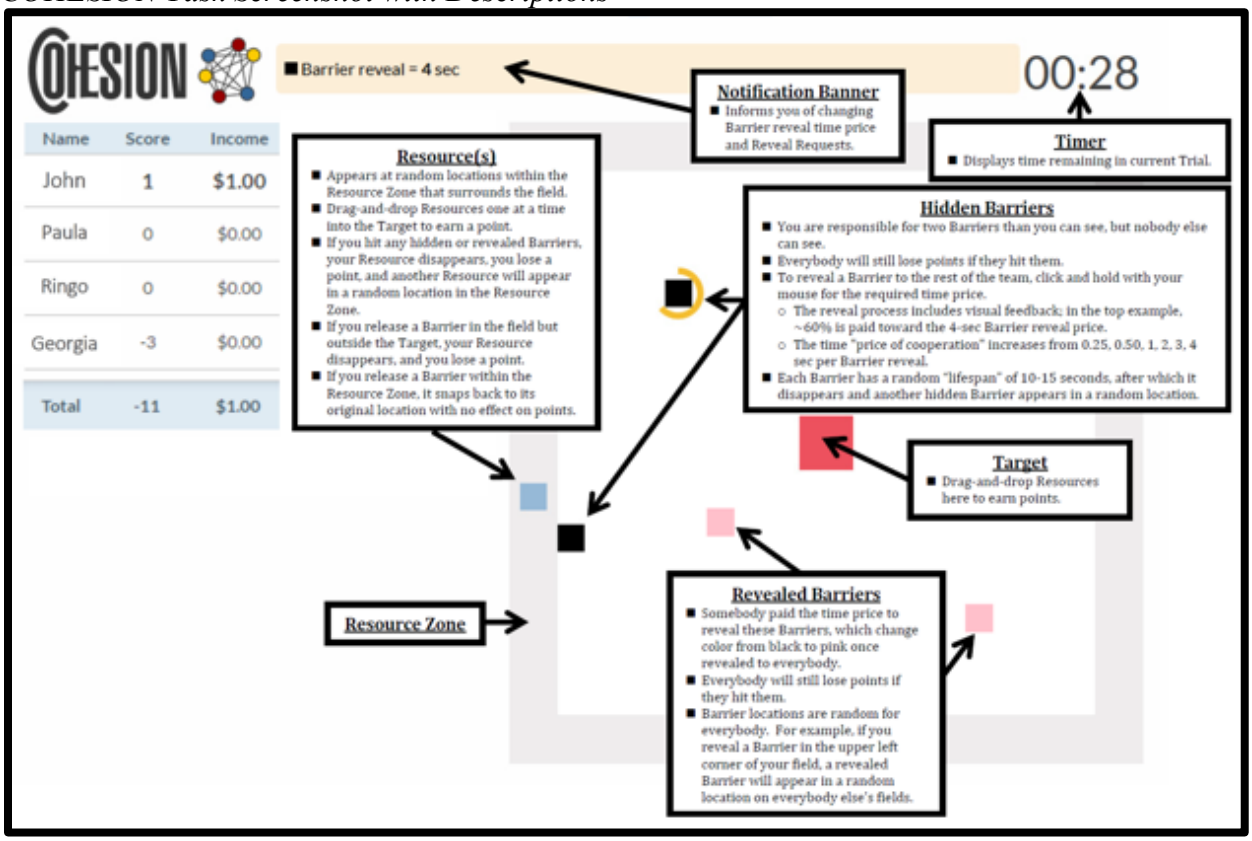
[1]To clarify, teams played for 45 minutes at a time, regardless of if (or how many times) they reached a game-over state.

COHESION

Developed and validated through NASA support, we used the COHESION task (Figure 5) as a standardized, objective, team-level measure of cooperative behavior (Hursh & Roma, 2013). COHESION is sensitive to personality, team composition, and external incentives, and correlates with social conflict and physiological stress (Roma et al., 2016; 2017).

Figure 5

COHESION Task Screenshot with Descriptions



The basic idea behind the game is that a team of three players attempts to gain rewards by moving "resources" into a central target area. Interfering with a player's performance is the presence of both revealed and hidden barriers between the resource area and the target area. Each player sees a different set of revealed barriers, which they can make visible to their teammates. The amount of time that must be invested in order to reveal a barrier increases every 30 seconds over the course of a 3-minute game. COHESION treats barrier reveals as cooperative behaviors and point scoring as selfish behaviors. It then calculates a fairness index for each 30-second segment, as well as each 3-minute game, based on the relative balance between the players of selfish vs. cooperative behaviors.

Our original intention was to use this fairness index as our primary metric of team cohesion, based on the hypothesis that teams with higher cohesion would have a better balance

of cooperative and selfish behaviors. As with CubeCrusher, our ability to use the fairness index was mitigated by the fact that teams would discuss ahead of time different strategies to experiment with. Some of these strategies involved one or two players only doing barrier reveals while the other player(s) scored points. That leads to a low fairness score even while the team is displaying high task cohesion. Other strategies allowed players to choose what actions they were going to perform, further confounding our ability to use fairness as a proxy for task cohesion.

Our solution to this is described in more detail below. It involved using clustering methods on a variety of COHESION metrics to identify sets of individual and team strategies that were being used. Then, instead of building models to predict the fairness index, we built models that used the physiological data from the Empatica to identify points when players changed their strategies, possibly indicating a breakdown in task cohesion if a player was deviating from the previously agreed upon team strategy.

Surveys and Journals

As part of the MEDULLA experiment, we also collected a variety of survey data (see Figure 3), as well as asked each participant to write a journal entry every night about that day's experiences. A full analysis of these surveys and journal entries was outside of the scope of CONCORDIA, but we were able to use an overview of these entries to, for example, identify that teams were counteracting the built-in metrics in CubeCrusher and COHESION by experimenting with different strategies.

Physiological Synchronization and Alignment

We used the Empatica E4 wristband device (https://www.empatica.com/research/e4/) to collect continuous physiological data, including activity/movement (via 3D accelerometer), HR, HRV, and EDA. Participants wore the device continuously throughout the study, including pre-mission training days, in-mission, and post-mission, with the exception of a one-hour charge and synchronization period every day.

Physiological synchronization was measured using the algorithm developed by Guastello and Peressini (2017). This algorithm assesses synchronization as a function of the coefficients generated by pairwise linear regression equations. We extended this algorithm to accept multivariable input vectors and determine the overall physiological synchronization of both IBI and EDA. We then calculated this synchronization for every COHESION segment and CubeCrusher game to serve as an input into our team cohesion classification systems as described below.

Team Strategy Identification and Deviance Detection

We pursued a methodology to detect and signal fluctuations in team task cohesion via the prediction of strategic dissent. This methodology may be split into three sections: recognizing behavioral strategies, systematizing physiologic signals, and predicting strategic deviations. We utilize the COHESION game as our surrogate for determining team cohesion. Specifically, we monitored trends in cooperative and selfish behavior allocations over the course of a trial. Our

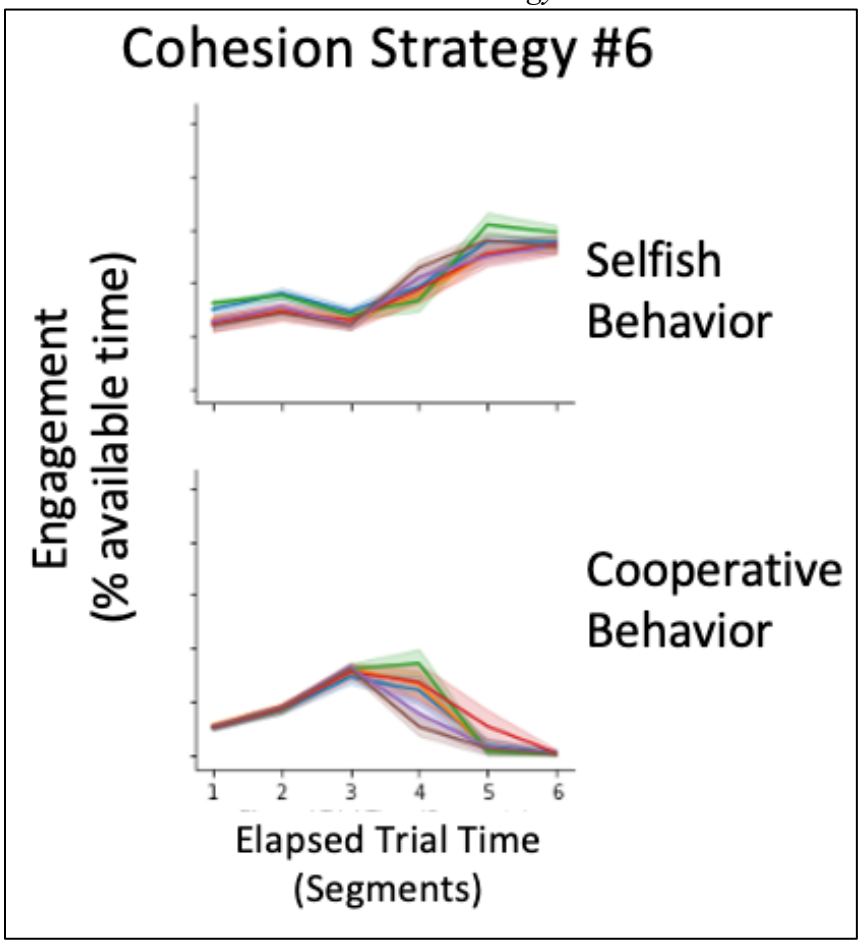
targeted metrics were selfish and cooperative engagement at the individual level and selfish and cooperative indices at the group level. Values were calculated per segment.

Strategy Detection

We define a "strategy" as any distinct behavioral allocation with shared, cross-segment, trends. Before diving into the technical details of strategy detection we present an example of an individual-level strategy to illustrate the concept. Figure 6 presents all the individual-level behavioral allocation trends classified by our model as belonging to Cohesion Strategy #6. Visualizing these trends together helps communicate what we characterize as a "strategy." We can see how individuals employing Cohesion Strategy #6 initially pursue both cooperative and selfish behaviors in fairly consistent allocations—cooperative engagement increasing as the price of cooperation increases from segment to segment—until segments 3 or 4, at which point they deem the opportunity cost for cooperation to be too high and dissent in favor of a more selfish behavioral allocation. Our approach to detecting these strategies can be broken down into two central steps: dimensionality reduction and clustering.

Figure 6

Individual-level COHESION Trial Strategy #6



Note. The x-axis reports elapsed trial time measured in segments. The y-axis records behavioral engagement measured as the percent of available time in each segment an individual spent pursuing a specific behavior.

In essence, our strategy detection system functions by clustering coordinate pairs representing an individual or group's behavioral allocation over the course of a trial. "Strategies" are defined by the resulting clusters. Our primary step during strategy detection is the derivation of coordinate pairs delineating cross-trial behavioral allocation. First, conceive an individual (or group's) behavioral allocation for a certain trial as a combination of two vectors: $\vec{s} =< s_1, s_2, s_3, s_4, s_5, s_6 >$ and $\vec{c} =< c_1, c_2, c_3, c_4, c_5, c_6 >$, each indicating the selfish (s) or cooperative (c) engagement over each segment within that trial. The coordinate pairs we use for clustering can be understood as $< s_t, c_t >$ where s_t is a value encompassing $\vec{s}$ and c_t is a value encompassing $\vec{c}$.

To reduce the dimensionality of behavior vectors to single values we use a technique called Uniform Manifold Approximation and Projection (UMAP). The UMAP is a dimensionality reduction technique that uses a combination of techniques from Riemannian geometry and algebraic topology to preserve the global structure of the data while reducing its dimensionality. It reduces high-dimensional data into a low-dimensional space while maintaining the similarities between data points. We can therefore take $S = \{\vec{s} =< s_1, s_2, s_3, s_4, s_5, s_6 >\in \mathbb{R}^6\}$ as the set of all selfish vectors and $C = \{\vec{c} =< c_1, c_2, c_3, c_4, c_5, c_6 >\in \mathbb{R}^6\}$ as the set of all cooperative vectors and approximate M_s and M_c—the approximate manifold structures of the vector-spaces S and C. We can then project each vector $\vec{s}$ and $\vec{c}$ onto a single dimension that best approximates the topological structure of their corresponding manifolds. This projection gives us a single value for each behavioral vector: s_t for $\vec{s}$ and c_t for $\vec{\vec{c}}$. We then combine these values into a single vector $\vec{b} =< s_t, c_t >$ delineating cross-trial behavioral allocation. This gives us a set $B = \{\vec{b} =< s_t, c_t >\in \mathbb{R}^2\}$ of all behavioral allocation trends utilized during a COHESION trial. We then pass B through to our clustering system.

To cluster our set of behavioral vectors (B) we utilize a technique called Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). HDBSCAN is a density-based clustering algorithm that groups nearby points into clusters. HDBSCAN identifies a dense region of points and marks it as a cluster. It then grows that cluster by adding nearby points with similar densities until the density drops below a specified threshold. Unlike other clustering algorithms, The HDBSCAN does not require pre-specification of the number of clusters. Moreover, it identifies and removes points that it classifies as "noise" or not-belonging to any cluster. This makes it highly suited for exploratory applications such as strategy detection. Furthermore, its density-based clustering mechanism is well suited to work in tandem with UMAP-reduced vectors designed to reduce dimensionality while preserving topological structure. Our clustering system clustered 99.3% of individual cross-trial behavioral allocations into one of 10 individual-level strategies and 94.1% of group cross-trial behavioral allocations into five group-level strategies (Figure 7).[2]

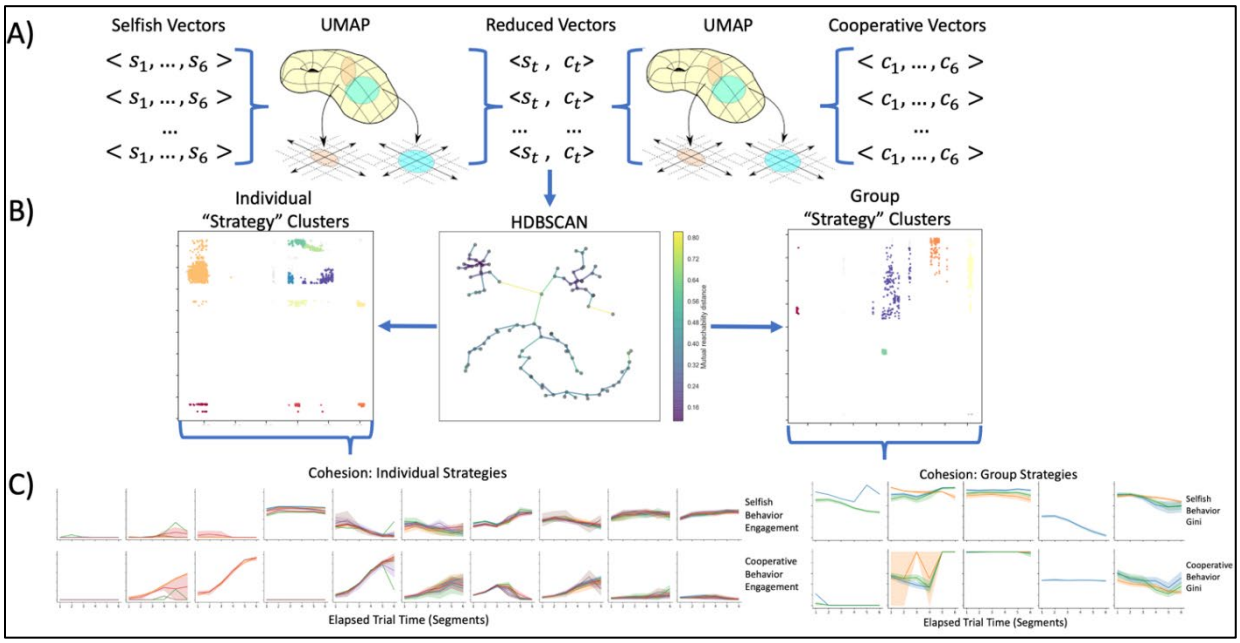
[2] Those not clustered were classified as noise.

Physiological Signal Extraction

Our pipeline for systematizing physiologic data signals restructures data from the Empatica E4 wristbands to use as standardized predictor variables in modeling strategic dissent. We target five of the six physiologic metrics measured by the Empatica wristbands: EDA, HR, BVP, IBI, and skin temperature. [3]

Figure 7

Illustrated Strategy Detection System



Note. A) Behavioral coordinate pair derivation via UMAP dimensionality reduction. B) Strategy clustering via HDBSCAN. C) Resulting COHESION trial strategies.

As these metrics were all recorded at different rates and were not aligned across all wristbands, data standardization was necessary to prepare the physiologic signals for modeling. The following processes were replicated for each metric:

1. We extracted the physiologic timeseries data corresponding to every COHESION trial.

2. We z-normalized these timeseries to focus model attention timeseries shape, rather than the absolute differences between timeseries.

3. We calculated pairwise distances between all the timeseries samples using Dynamic Time Warping (DTW). The DTW is a technique used in the timeseries analysis for comparing and aligning two sequences, even if they have different lengths and rates of progression. It calculates the minimum cumulative distance between points of the two sequences and warps the

[3] As participants were stationary during COHESION sessions the accelerometer data from the Empatica watches were omitted from our pipeline.

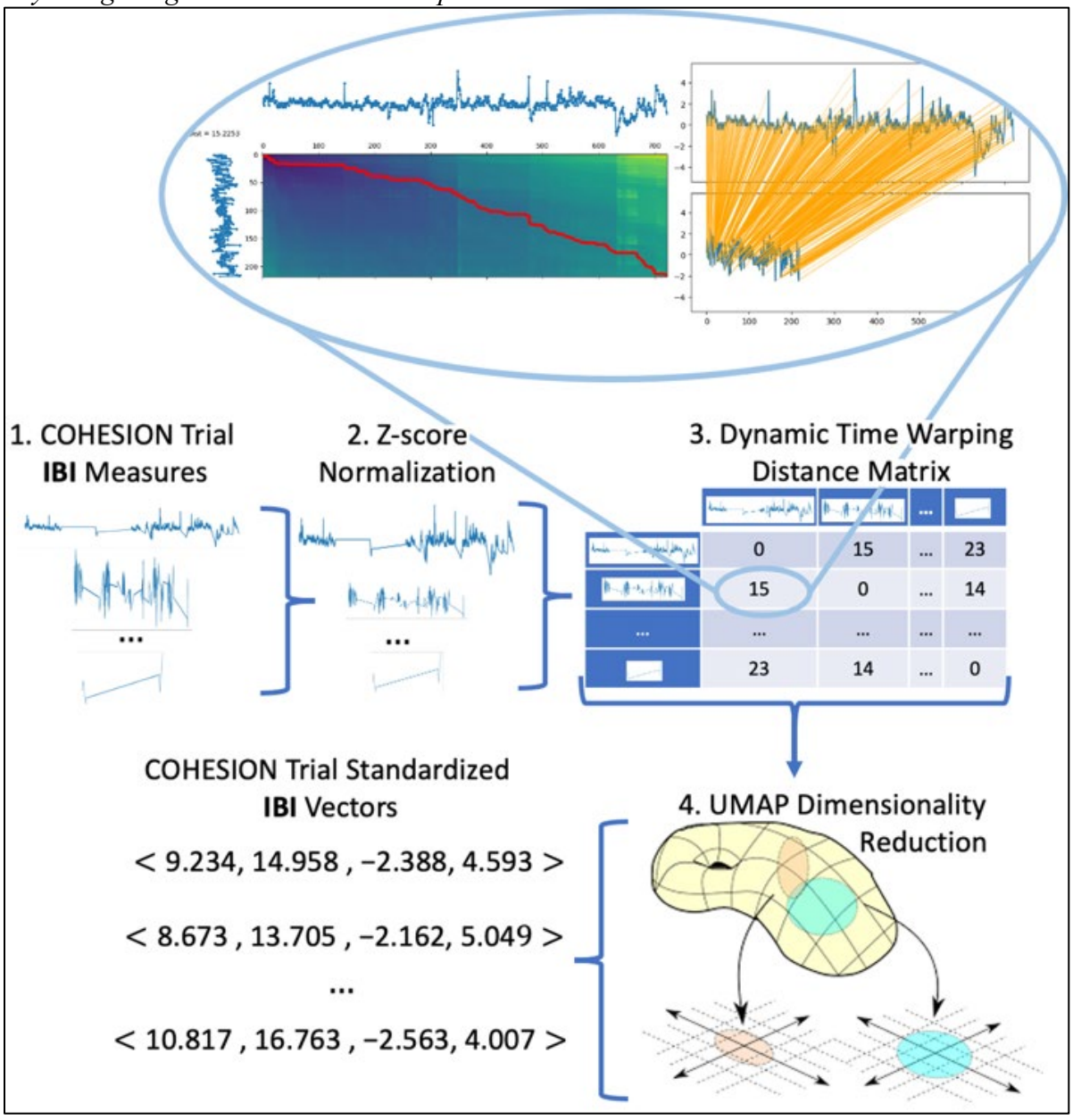
time axis non-linearly to align them. This makes DTW useful for comparing sequences that have similar patterns but are not the same.

4. We used UMAP over the precomputed distance matrix to reduce all timeseries vectors to four dimensions.

Figure 8 illustrates this process via the IBI signal. Executing this system across all physiologic measures outputs a set of five vectors—one for each physiologic metric of interest—each vector comprised of four values communicating their position in relative space to one another.

Figure 8

Physiologic Signal Standardization Pipeline Illustrated Over the Interbeat Interval Metric



Predicting Strategic Dissent

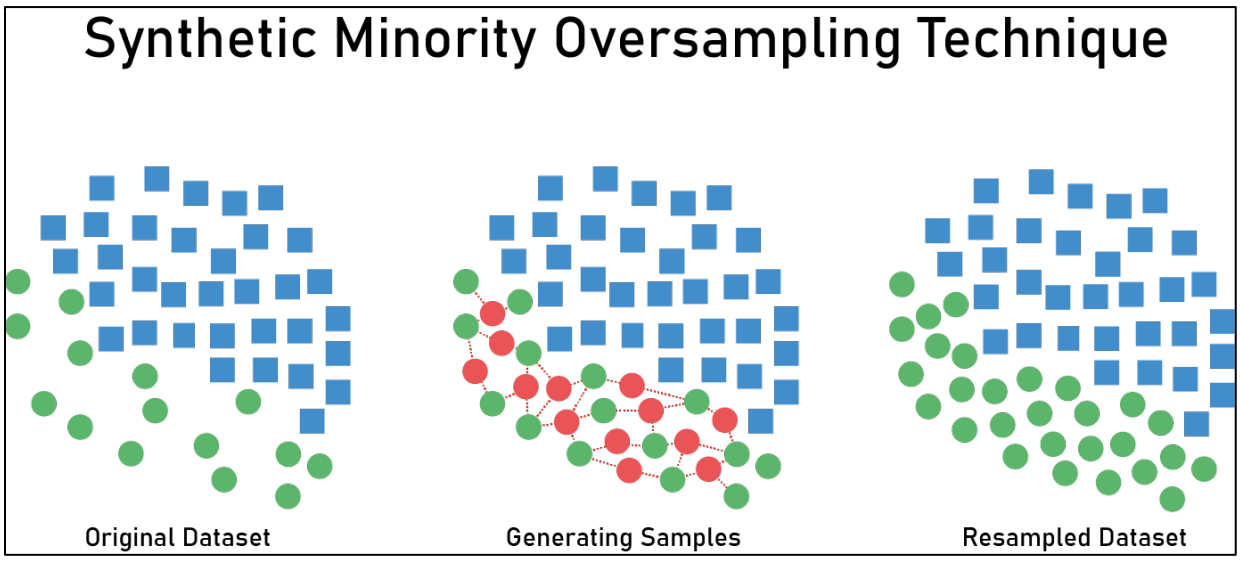
During our predictive modeling phase, we constructed two logistic regression models. Our first model, $Dissent_{pgms(t)} = Empatica_{pgms(t)} + Empatica_{pgms(t-1)} + Dissent_{pgms(t-1)}$, predicted whether individual p from mission m, group g, would dissent from their current strategy in session s trial t ($Dissent_{pgms(t)}$) given their current standardized Empatica data ($Empatica_{pgms(t)}$), their standardized Empatica data from the previous trial ($Empatica_{pgms(t-1)}$), and whether or not they changed strategies in the previous trial ($Dissent_{pgms(t-1)}$).

Our second model, $Dissent_{gms(t)} = \sum_{p=1}^{|g|} Dissent_{pgms(t)} + Dissent_{gms(t-1)}$, predicted whether group g from mission m, would dissent from their current strategy in session s trial t ($Dissent_{gms(t)}$) given how many team members were predicted to switch strategies in the same trial ($\sum_{p=1}^{|g|} Dissent_{pgms(t)}$) and whether or not the group changed strategies in the previous trial ($Dissent_{gms(t-1)}$).

Because strategic consistency was more common than strategic dissent, we had to address the label imbalance present in our data before we could train our models to their targeted tasks. To do this we employed a technique called Synthetic Minority Oversampling Technique (SMOTE). The SMOTE (Figure 9) generates new synthetic samples for the minority class (dissent) instead of simply duplicating existing samples. The new samples are generated by taking the feature-wise difference between two randomly selected minority class samples and adding it to a randomly selected sample from the same class to create a new, synthetic sample. This technique helps balance the class distribution by generating new, synthetic samples for the minority class, leading to improved performance in models trained on imbalanced datasets.

Figure 9

SMOTE Sample Generation Visualization



After addressing the target label imbalance, we trained each of our models employing 5-fold cross-validation to avoid overfitting (Figure 10) and tuned a set of six hyperparameters over the validation sets, maximizing model accuracy. The hyperparameters tuned were: (1) the minority class re-sampling rate used during SMOTE; (2) the solver algorithm used in the optimization; (3) a numeric penalty designed to discourage high model complexity thereby reducing model generalization error and regulating overfitting; (4) the regularization strength or the relative importance weight is given to training data vs. validation data when fitting; (5) model tolerance, which stops searching for optimality once a certain metric threshold is achieved (in our case model accuracy); and (6) the maximum number of optimization iterations for the tuning algorithm to attempt. Table 1 shows the final optimal hyperparameters for training our logistic regressions.

Figure 10

5-Fold Cross-Validation Visualized

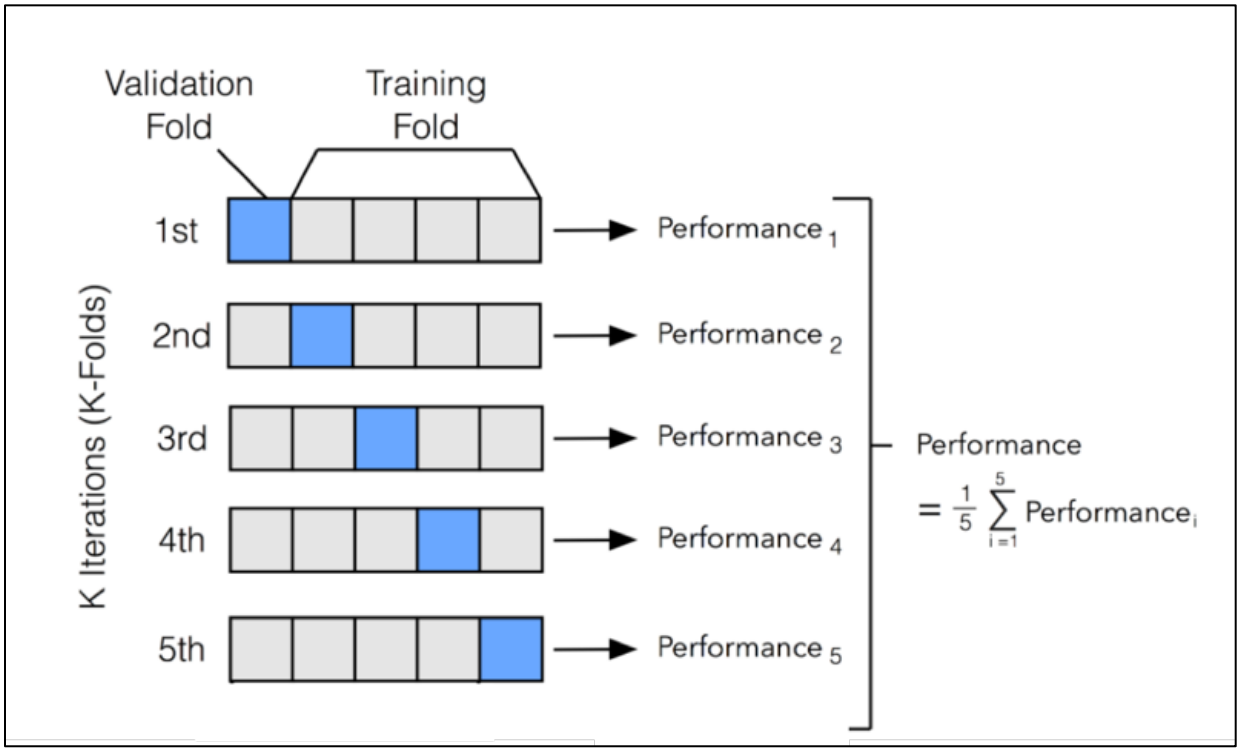


Table 1

Hyperparameter Tuning Results for Individual (Model 1) and Group (Model 2) Dissent Prediction Logistic Regression Models

	SMOTE Sampling Strategy	Solver	Penalty	C	Tolerance	Max Iter
Model 1 (Individual)	0.99	liblinear	12	7.02	0.002	483
Model 2 (Group)	0.93	liblinear	12	2.63	0.0002	385

Finally, after training and tuning our model, we calculated three probability thresholds to optimize precision, recall, and F1 score (Table 2). These thresholds were used to classify the logistic regression's outputs as one if the probability of dissent was above the threshold, and 0 if the probability of dissent was below the threshold. For our final predictions, we utilized the F1-optimized threshold. Results from applying each model to our test set are shown in Table 2. Figure 11 shows the corresponding confusion matrices for each model's performance at strategic dissent prediction.

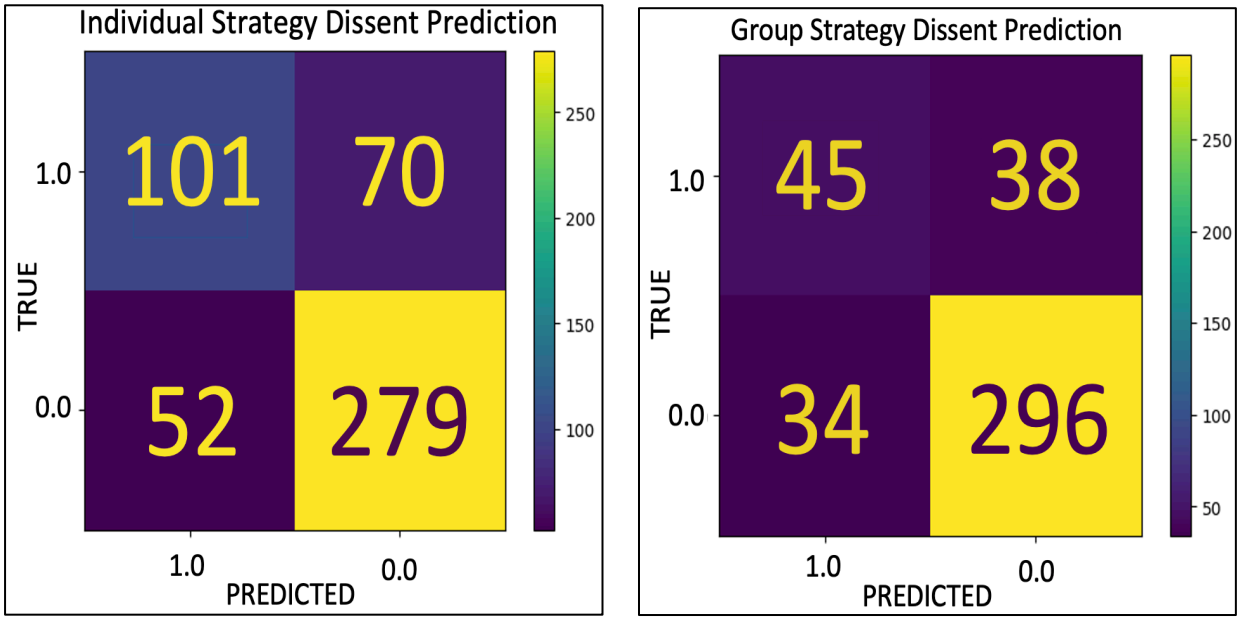
Table 2

Model Performance Results and Metric Optimized Thresholds for Logistic Regressions Models at Predicting Strategic Dissent

		Model	Accuracy	Prediction	Recall	F1
I	Model Performance	Individual Dissent	0.79	0.50	0.66	0.62
		Group Dissent	0.83	0.54	0.57	0.56
II	Optimized Thresholds	Individual Dissent		0.79	0.30	0.41
		Group Dissent		0.84	0.35	0.77

Figure 11

Confusion Matrices Showing Performance at Predicting Strategic Dissent

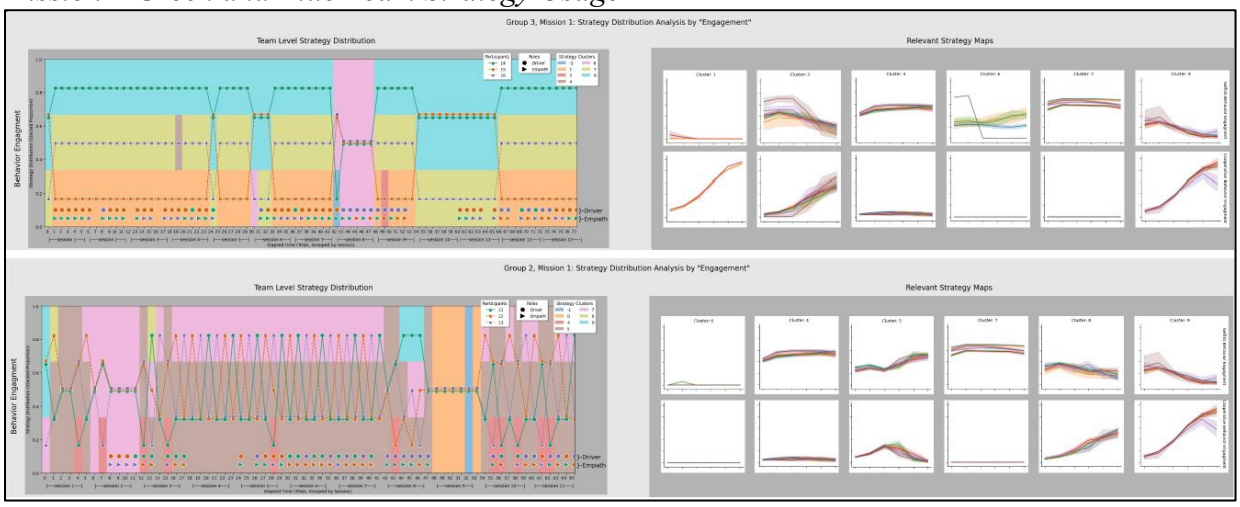


Note. Left: Model 1, Individual dissent prediction. Right: Model 2, group dissent prediction.

Figure 12 (top) shows the strategy usage for the Mission 1 Green team. Without worrying about what their specific strategies were, what should stand out is that they are extremely consistent in their strategy usage across almost the entire mission. Figure 12 (bottom) shows the strategy usage of the Mission 1 Blue team. While the team seemed to have a consistent set of strategies that they stuck to, individuals were highly variable as to which strategies they were using during any particular game or segment.

Figure 12

Mission 1 Green and Blue Team Strategy Usage



Note. Top: Green team strategy usage. Bottom: Blue team strategy usage.

Based on these results, our hypothesis would indicate that the M1 Green team would have higher team cohesion than the M1 Blue team. However, the journals and surveys from Mission 1 indicate the exact opposite. In fact, before running these analyses our subjective examination led us to believe that the M1 Green team had the lowest team cohesion of any team across all of our missions, while the M1 Blue team was among the most cohesive teams. M1 Green had one team member in particular who would not communicate or cooperate with their teammates, which led to a lot of stress and frustration.

One possible explanation for the counterintuitive results could be that the M1 Green team did not plan out different potential strategies to maximize their performance scores. The M1 Blue team displayed a wider variety of strategy usage precisely because they decided as a team to investigate different strategies. Thus, in this context, consistent strategy usage indicates lower team cohesion rather than higher.

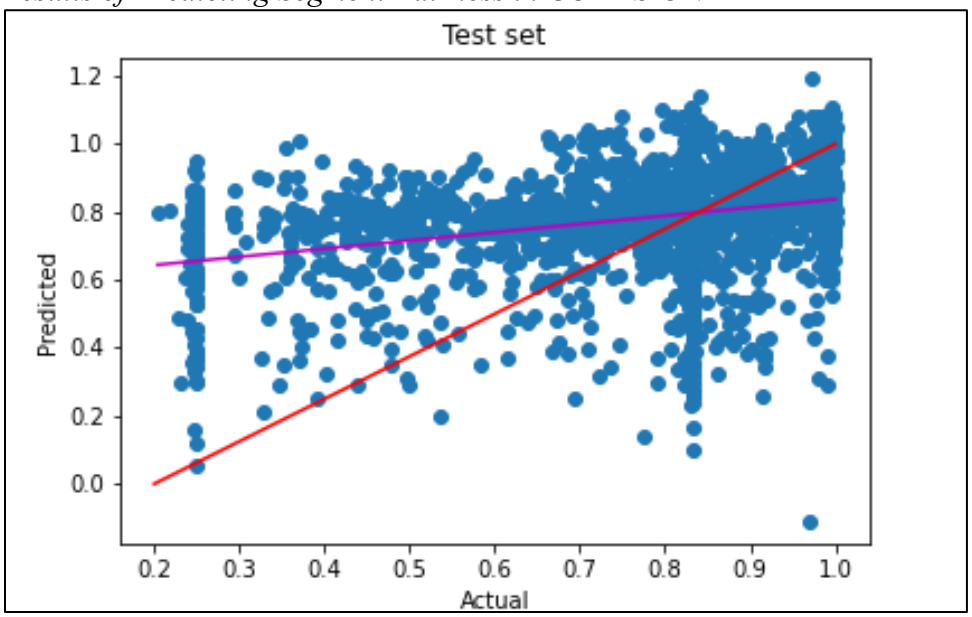
Performance Predictions

We also developed models for predicting performance on COHESION and CubeCrusher. Specifically, we wanted to predict the segment-fairness-index in COHESION and metrics in CubeCrusher related to multiple people participating in completing columns. For COHESION, we constructed a three-layer long short-term memory (LSTM) network that took as inputs values from all six segments of a single COHESION game. There were seven inputs: mean IBI for each of the three players, mean EDA for each player, and the physiological synchronization metric described above.

Figure 13 shows the results from this model. The model predicts better than chance and finds the central tendency of the distribution of performance scores. However, there is still room for improvement.

Figure 13

Results of Predicting Segment Fairness in COHESION



Note. Red line is a perfect result and purple line is the best fit line.

In trying to understand how the model could have performed better, we looked at the relationship of each of the input variables to the segment-fairness-index. Figure 14 shows that none of the input variables has a very strong relationship with the segment-fairness-index. Thus, it may well be the case that the model presented has captured as much of the predictive value of these variables as can be captured.

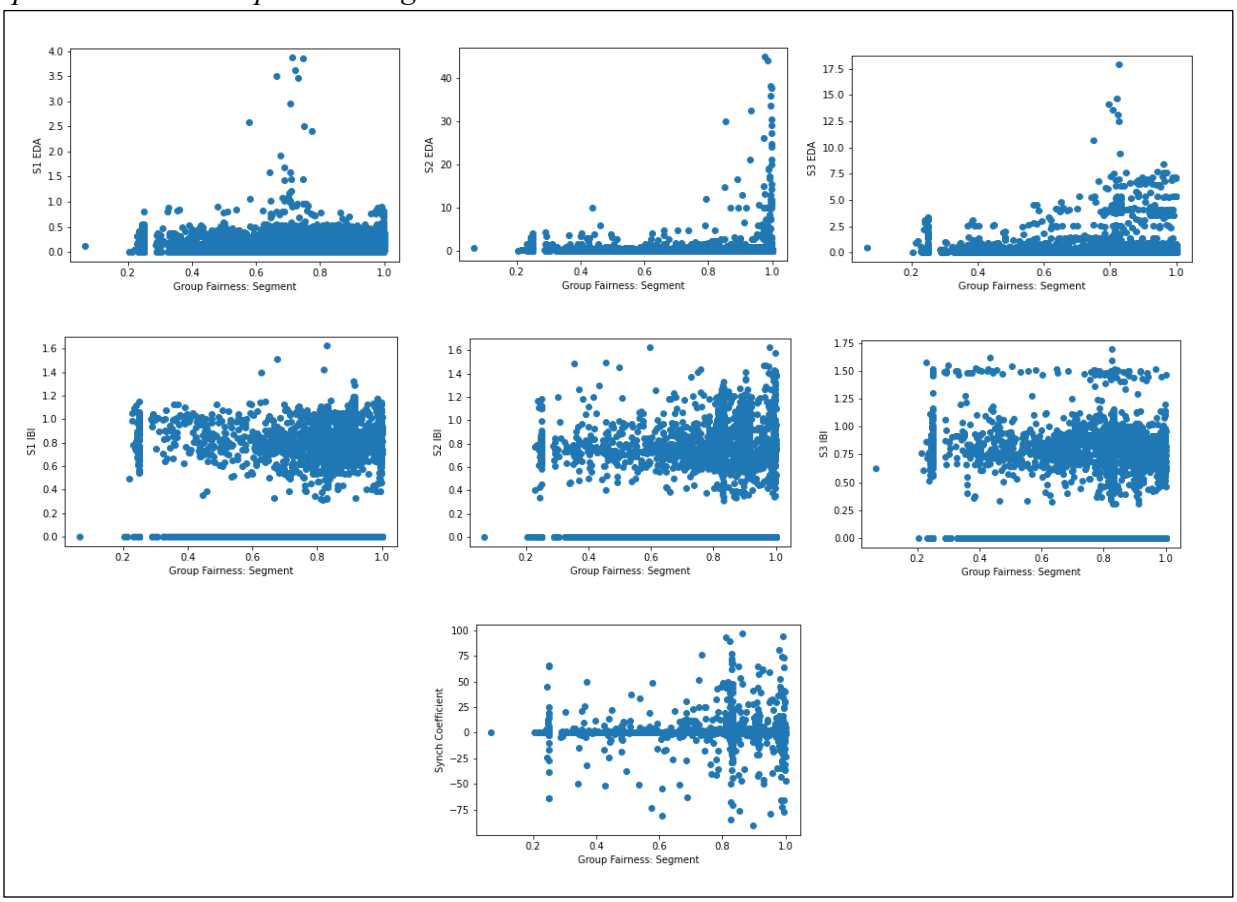
We constructed a similar model to predict the CubeCrusher metrics of the number of columns finished using pieces from all three players and also the average number of players contributing pieces to all the finished columns in a given CubeCrusher game. As can be seen in Figure 15, the results were similar to those from COHESION.

Linguistic Synchronization

The primary aim of CONCORDIA is to examine the feasibility of leveraging physiological metrics to measure team cohesion. However, we also collected communication data during COHESION and CubeCrusher, automatically transcribed from video recordings using Google Cloud's ‘Speech-to-Text’ application programming interface. Between our four missions, this produced a total of 66,931 utterances made in conversation by our 24 participants. This provided an opportunity to examine linguistic synchronization as an exploratory consideration. Ultimately, we are interested in determining the extent to which linguistic synchronization provides valuable information about team cohesion. Similar to our overall expectations, we expected that teams with higher linguistic synchronization would also display higher team cohesion.

Figure 14

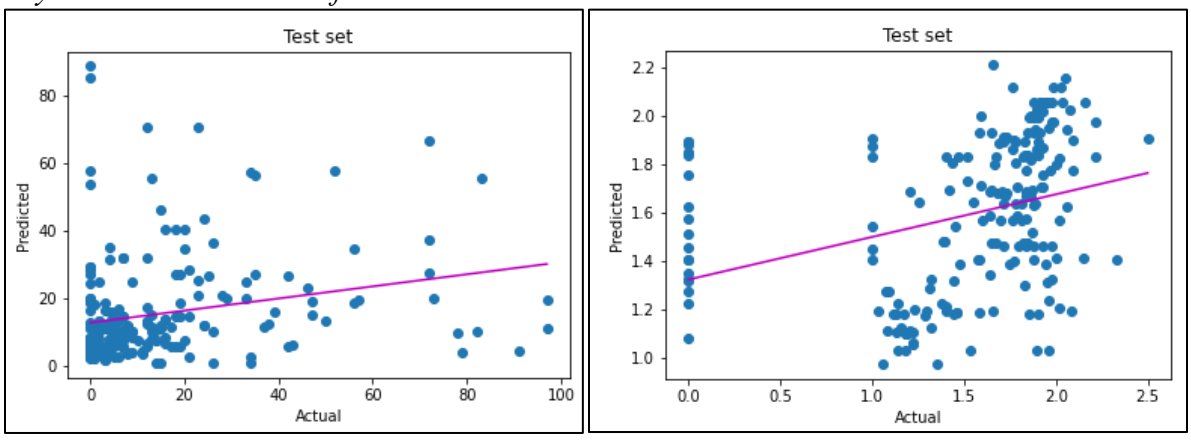
Input Variables Compared to Segment-Fairness-Index



Note. None show an especially strong relationship.

Figure 15

Physiometric Predictions for CubeCrusher



Note. Left: Number of columns that all three players contributed pieces to. Right: Average number of players who contributed to the columns in a game.

To characterize our transcripts, we used a series of fine-tuned large language models. Publicly available, pre-trained, large language models leverage contextual information from their pretraining datasets to achieve high performance on downstream tasks they can be fine-tuned to perform. We use BERT-based (Bidirectional Encoder Representations from Transformers) models from the HuggingFace library for two tasks: data anonymization and sentiment analysis.

Transcripts were anonymized using a BERT model fine-tuned on the task of recognizing spans of text that represent characters. This model architecture consists of a single linear layer after the hidden-states output of the BERT transformer. This model's training dataset consists of hand-annotated data by our in-house researchers, achieving an F1 score of 0.94. Since it is possible to do a name-and-replace search of participants' names, this automatic method protects identification of characters mentioned outside the scope of the project (e.g., scientists' names, participants' family members).

Our models for sentiment analysis are fine-tuned on the task of sentiment intensity regression. Our model architecture consists of a BERT-base-uncased model, a pooling function, and a single linear neural layer. We created multiple models to classify our transcripts along several features: anger, fear, joy, sadness, and politeness.

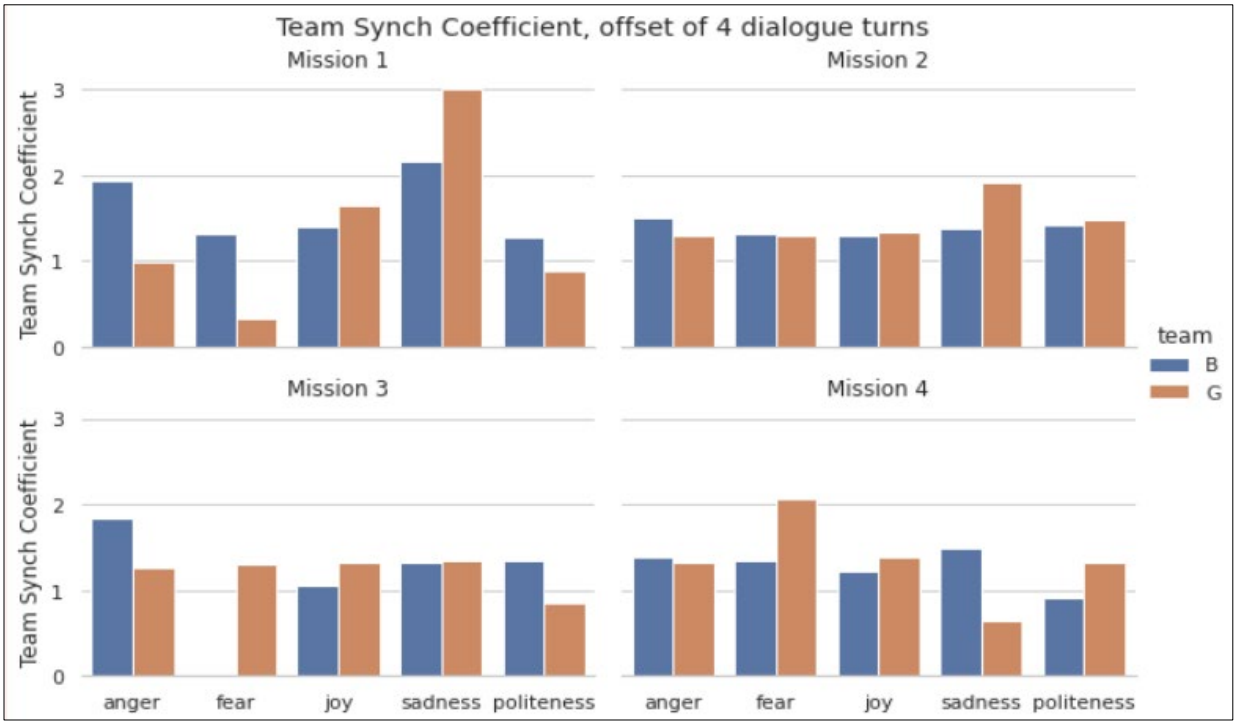
The dataset for anger, fear, joy, and sadness is sourced from SemEval2018 Task 1 which provides a training, evaluation, and test set for detecting emotion intensity along a regressive scale of 0 (low emotion) and 1 (high emotion). Our fine-tuning dataset for politeness is from PERFECTA, which created a hand-annotated dataset for classifying politeness along a regressive scale of -1 (rude) to 1 (polite). Our BERT models achieve a mean squared error within the range of 0.02-0.03 on their test datasets when fine-tuning the BERT-base-uncased model along 10 epochs. Our resulting models use Z-score normalization to keep information about outliers while maintaining the desired range for each feature.

We use synchronization metrics inspired by Guastello and Peressini (2016). Here, synchronization is the amount of influence participants may have with their conversation partners. This model is built on the concept of “empaths,” those who are most emotionally influenced, and “drivers,” those that influence trends of emotion in conversation. For our analysis, team synchronization models the amount of influence each participant has on each other participant on their team, relative to each of our five features of sentiment. For each one of our sentiment features, each team's synchronization metric is determined from a matrix of synchronization coefficients. Each value of the pairwise synchronization coefficient matrix is setup as a linear regression problem between every pair of participants on the same team. Our final synchronization matrix is normalized first with power transformation to produce a Gaussian shape. Next, min-max scaling within the range 0-3 is applied to maintain outliers and indicate “low,” “average,” “medium,” and “high” synchronization (see Figure 16).

In summary, the M1 Green team displayed the highest overall synchronization score. The prior results suggest that the M1 Green team also had the lowest cohesion of any of the mission teams. So, our exploratory approach to linguistic synchronization is clearly not an accurate reflection of cohesion; in fact, it appears to reflect the opposite.

Figure 16

Linguistic Syncronization for a Variety of Emotional Factors



Note. B = Blue team, G = Green team.

MSTC Field Test

Setup and Procedures

We were invited to use the Empatica devices to collect data from Soldiers participating in the Combat LifeSaver (CLS) course offered at the Fort Cavazos Medical Simulation Training Center (MSTC). Six Soldiers wore the devices all day during their class days. We also collected self-reported team cohesion and performance measures from the Soldiers, as well as similar measures of the team from the course instructors. The surveys are included in Appendix A and B.

Days 1-3 of the CLS course involved almost exclusively classroom instruction. While we collected data on these days, it was not directly relevant to the goal of assessing team cohesion in extreme environments. It also did not involve a lot of teamwork by the Soldiers. We kept that data as a potential baseline in future analyses, but otherwise did not use it for CONCORDIA. Day 4 tested the Soldiers' ability to work as a team implementing the procedures they had learned in the classroom. Soldiers completed a physically intensive obstacle course involving a simulated water crossing and casualty evacuations from a tank and helicopter.

Results

We input the data collected from Day 4 into our strategy deviation classifier described above. Our hypothesis was that any identified strategy deviations might indicate faltering team cohesion from the team member involved. We ran into three significant hurdles with validating this hypothesis.

First, three of the Empatica devices failed completely when they were submerged during a crawl through a muddy tunnel. As this occurred at the very beginning of the exercise, it meant that we gathered no data from those devices for the entire day.

Second, the surveys showed very little variation, both from the Soldiers and from the instructors. Full results are included in Appendix C, but essentially the Soldiers and instructors rated everything a 4 or a 5 on a five-point Likert scale.

Third, the nature of the exercise was such that we lacked the kind of precise timing of events and activities that the Soldiers were doing that we had for participants with COHESION and CubeCrusher. Thus, it is difficult to pinpoint whether any detected deviations corresponded to actual events during the day.

Compounding all these difficulties, our strategy deviation detection system failed to detect any strategy deviations. Given the ratings from the Soldiers and instructors, it is possible that there were no significant deviations during the day, at least among the three Soldiers we had data for. It could also be the case that our system simply failed to identify deviations that did occur. Without any positive identifications or a precise timeline of the day's events, it is impossible to determine which is the case.

WATCH Interface

A secondary and concurrent aspect of CONCORDIA involved the development of a user interface to display team cohesion metrics based on physiological data. Our aim was to offer practical value by developing a prototype interface for Army leaders to track the cohesion of their teams in real time.

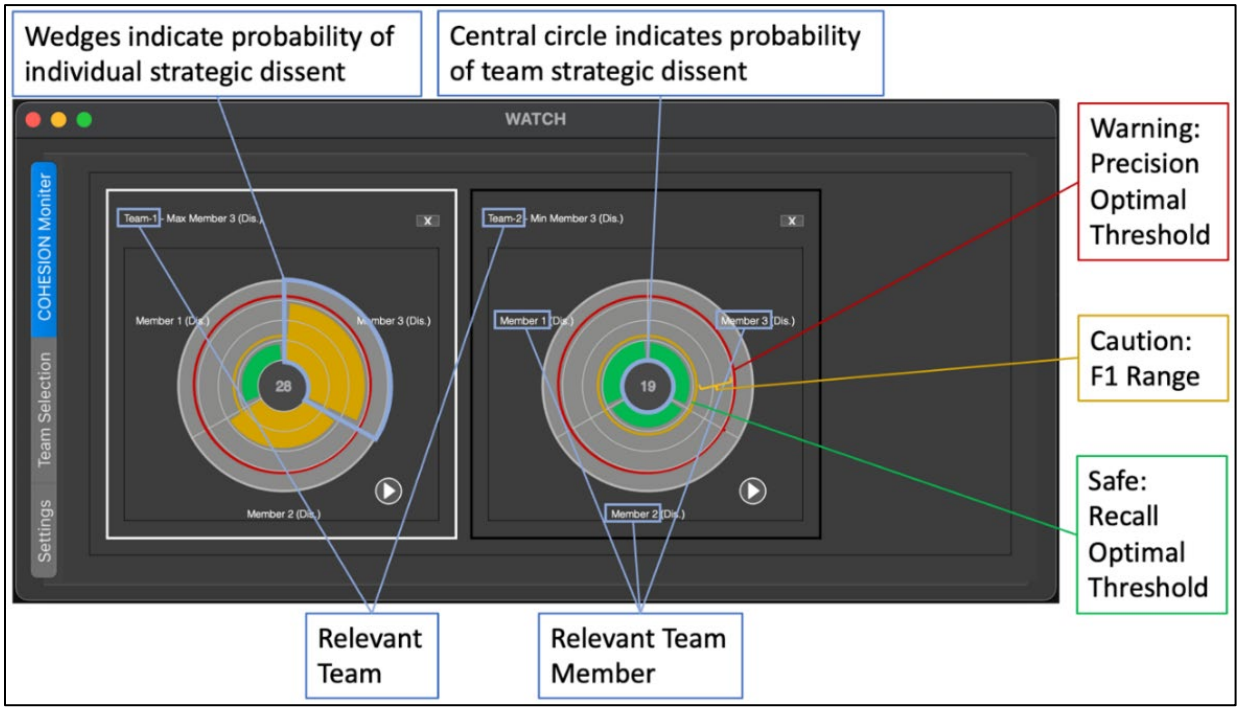
The Widget for Awareness of Team Cohesion Heuristics (WATCH) is our front-end user interface for monitoring team cohesion. The WATCH operates over a dynamic data stream, presenting real-time strategic dissent-predicted probabilities in stop-light, radial diagrams, designed to facilitate rapid comprehension of strategic dissent predictions. WATCH includes Team Selection and Settings panels to allow users to customize their central cohesion monitor view.

The primary *COHESION Monitor* tab (Figure 17) allows users to supervise the predicted task cohesion of selected teams through WATCH's radial, stop-light visualizations. Each visualization wedge projects each team member's predicted probability of strategic dissent. The central circle displays the predicted probability of team strategic dissent. Default thresholds for stop-light color indications are determined during the modeling phase. During threshold optimization we record the probability thresholds resulting in maximized recall and precision scores. High recall is achieved by low false consistency predictions and high precision is achieved by low false dissent predictions.

If the probability of an individual team member dissenting is below the recall optimized threshold it is highly *unlikely* that the individual will dissent, hence colored green. Conversely, if the probability of an individual team member dissenting is above the precision optimized threshold it is highly *likely* that the individual will dissent, therefore colored red. Any probability between the safe zone (recall optimized threshold) and warning zone (precision optimized threshold) is of moderate risk of dissent, and thereby colored yellow (Figure 17).

Figure 17

Screenshot of WATCH UI: COHESION Monitor Tab with Highlighted Features



The *Team Selection and Settings* tabs allow users to customize the central cohesion monitor. Under the *Team Selection* tab (Figure 18) users may quickly review all teams' probabilities of group dissent (Figure 18: Group P(dissent)) and select those that they wish to monitor more closely (Figure 18: Selection Panel). Selected changes may be saved or canceled by clicking the corresponding button above the tabular review panel. After applying changes, selected teams will appear on the *COHESION Monitor* tab.

The *Settings tab* (Figure 19) allows the user to customize the radial, stop-light, and visualizations under three central customization fields: metric selection, metric weights, and caution and warning thresholds. The "Show" container allows the user to select which team members they wish to monitor by checking the respective boxes (Figure 20: second row).

Figure 18

Screenshot of WATCH UI: Team Selection Tab with Highlighted Features

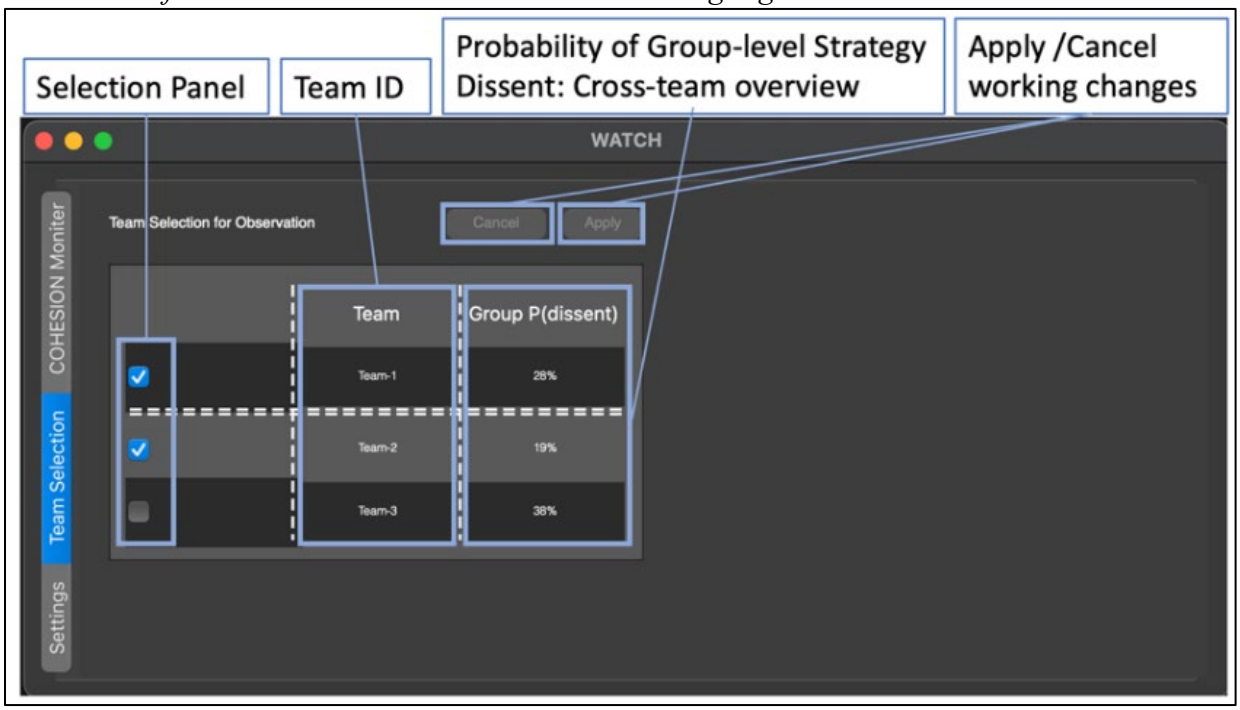
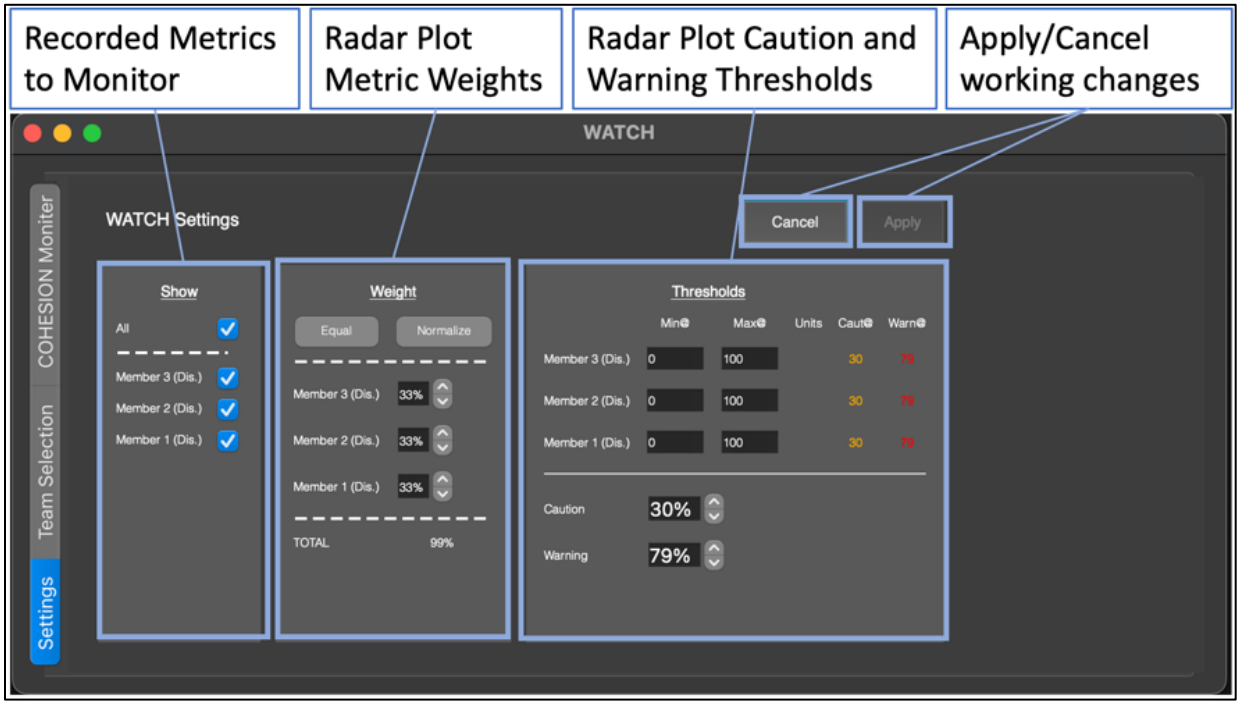


Figure 19

Screenshot of WATCH UI: Settings Tab with Highlighted Features



If instead the user wishes to focus primarily on specific team members but keep others under peripheral monitoring, they may alter the radial wedge weighting via the "Weight" container (Figure 20: third row). To do this, one would decrease the weighting of peripheral members and select "Normalize" to ensure total radial coverage. Corresponding visualization wedges would then be sized according to specified weighting. Selecting "Equal" will return wedges to equal weighting. The "Thresholds" container allows users to adjust the caution and warning thresholds. Users may adjust individual members' caution and warning thresholds by altering the minimum and maximum values in the top section of the "Thresholds" container (Figure 20: final row).

Local threshold changes do not alter the threshold markers but modify the multiplier sizing of each of the wedges. In this way, although all team members may have different numeric thresholds for caution and warning, they may be monitored on the same radial scale, for comparative interpretation. Global caution and warning thresholds may be adjusted by increasing or decreasing the global percent-based thresholds in the lower section of the "Thresholds" container. These changes can be reset at any time. As with the *Team Selection* tab, all working changes must be activated by selecting the "Apply" button or canceled by selecting the "Cancel" button.

Conclusions

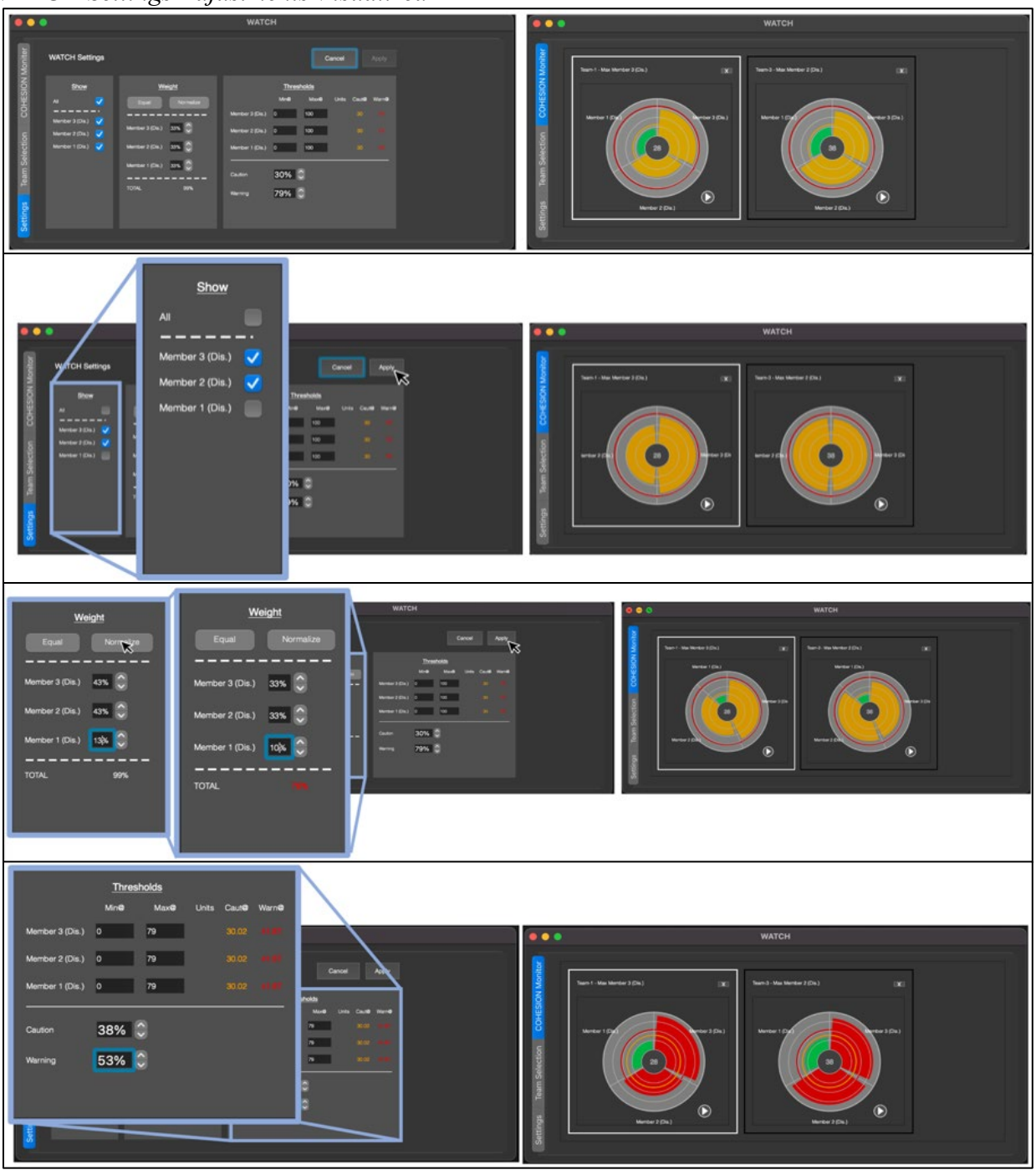
We believe that CONCORDIA represents a demonstration of the feasibility of making at least broad assessments of team cohesion using the data from wearable devices and presenting these data in a user-friendly interface. Through the course of this project, we identified future directions in which to carry this research forward so as to improve on the results presented here.

First among these is the need to either identify or develop better longitudinal metrics for team cohesion. The metrics built into both COHESION and CubeCrusher failed to capture the range of possible behaviors indicating high team cohesion. This was largely because team members had the opportunity to discuss game strategies they were going to adopt as a team, and essentially conduct mini-experiments during their game play to assess which strategies led to the highest level of performance. The need for better longitudinal metrics of team cohesion has also been identified by, among others, Salas et al. (2015).

One way to accomplish the above might be to develop methods to account for longitudinal context in existing metrics. For example, the segment-sharing-index was not an ideal metric in part because teams adopted strategies that were explicitly unfair to individuals in the hope that such strategies would lead to better overall team performance. A common strategy involved cooperative behaviors for one or two players during the early game segments when such behaviors are relatively cheap, combined with selfish behaviors for the remaining player(s). That strategy is associated with a very low segment-fairness-index, yet team cohesion is high to the extent that all the team members persist with said strategy. That said, our own strategy identification system indicated that merely maintaining a consistent strategy is not, in and of itself, indicative of high team cohesion. Thus, our second recommendation for future work involves trying to determine ways to quantify the effects of context on these types of task-based team cohesion assessments.

Figure 20

WATCH Settings Adjustments Visualized



Our third recommendation for future research is to develop a more controlled type of field assessment. At the least, any future field test of CONCORDIA, especially in a live training environment, should include a more fine-grained characterization of exactly what the participants are doing at any specific point in time so that changes in physiological data can be associated

with specific events. This will allow for a more robust association of the physiological metrics with behaviors that contribute to team cohesion and in turn performance.

Finally, the development of an interface to display the team cohesion metrics in real time is a promising practical application of CONCORDIA. Our user interface was designed to be readily interpretable and intuitive, based on stop-light, radial diagrams. The ultimate end-users of this interface are Army leaders operating in extreme environments. However, WATCH is a prototype because it has not been implemented in an operational context, nor has it been evaluated in any systematic way. Thus, in addition to follow-on field evaluations, another direction for future research is to conduct a user evaluation study to ensure that the WATCH interface is useable in an Army context (e.g., Brooke, 1996).

References

Appelhans, B. M., & Luecken, L. J. (2006). Heart rate variability as an index of regulated emotional responding. *Review of General Psychology*, *10*(3), 229-240. https://doi.org/10.1037/1089-2680.10.3.229

Baker, D. P., & Salas, E. (1997). Principles and measuring teamwork: A summary and look toward the future. In M. T. Brannick, E. Salas, & C. Prince (Eds.), *Team performance assessment and measurement: Theory, methods, and applications* (pp. 331–355). Erlbaum.

Beal, D. J., Cohen, R. R., Burke, M. J., & McLendon, C. L. (2003). Cohesion and performance in groups: a meta-analytic clarification of construct relations. *Journal of Applied Psychology*, *88*(6), 989-1004. https://doi.org/10.1037/0021-9010.88.6.98

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P.W. Jordan, B. Thomas, B. A. Weerdmeester & I. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189-194). Taylor & Francis.

Castaño, N., Watts, T., & Tekleab, A. G. (2013). A reexamination of the cohesion–performance relationship meta-analyses: A comprehensive approach. *Group Dynamics: Theory, Research, and Practice*, *17*(4), 207-231. http://dx.doi.org/10.1037/a0034142

Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., & Calantone, R. (2017). The promise and perils of wearable sensors in organizational research. *Organizational Research Methods*, *20*(1), 3-31. https://doi.org/10.1177/1094428115617004

Chiocchio, F., & Essiembre, H. (2009). Cohesion and performance: A meta-analytic review of disparities between project teams, production teams, and service teams. *Small Group Research*, *40*(4), 382-420. https://doi.org/10.1177/1046496409335103

Driskell, T., Salas, E., & Driskell, J. E. (2018). Teams in extreme environments: Alterations in team development and teamwork. *Human Resource Management Review*, *28*(4), 434-449. https://doi.org/10.1016/j.hrmr.2017.01.002

Evans, C. R., & Dion, K. L. (1991). Group cohesion and performance: A meta-analysis. *Small Group Research*, *22*(2), 175-186. https://doi.org/10.1177/1046496491222002

Grossman, R., Nolan, K., Rosch, Z., Mazer, D., & Salas, E. (2022). The team cohesion-performance relationship: A meta-analysis exploring measurement approaches and the changing team landscape. *Organizational Psychology Review*, *12*(2), 181-238. https://doi.org/10.1177/20413866211041157

Guastello, S. J. (2016). Physiological synchronization in a vigilance dual task. *Nonlinear Dynamics, Psychology, and Life Sciences*, *20*(1), 49–80. https://www.societyforchaostheory.org/ndpls/askFILE.cgi?vol=20&iss=01&art=03&desc=PDF

Guastello, S. J. (2017). Nonlinear dynamical systems for theory and research in ergonomics. *Ergonomics*, *60*(2), 167–193. https://doi.org/10.1080/00140139.2016.1162851

Guastello, S. J., & Peressini, A. F. (2017). Development of a synchronization coefficient for biosocial interactions in groups and teams. *Small Group Research*, *48*(1), 3–33. https://doi.org/10.1177/1046496416675225

Gully, S. M., Devine, D. J., & Whitney, D. J. (1995). A meta-analysis of cohesion and performance: Effects of level of analysis and task interdependence. *Small Group Research*, *26*(4), 497-520. https://doi.org/10.1177/1046496495264003

Hursh, S. R., & Roma, P. G. (2013). Behavioral economics and empirical public policy. *Journal of the Experimental Analysis of Behavior*, *99*(1), 98-124. https://doi.org/10.1002/jeab.7

Kazi, S., Khaleghzadegan, S., Dinh, J. V., Shelhamer, M. J., Sapirstein, A., Goeddel, L. A., Chime, N. O., Salas, E., & Rosen, M. A. (2021). Team physiological dynamics: A critical review. *Human Factors*, *63*(1), 32-65. https://doi.org/10.1177/0018720819874160

Kozlowski, S. W. J., Chao, G. T., Chang, C.-H., Fernandez, R. (2016). Team dynamics: Using "big data" to advance the science of team effectiveness. In Tonidandel, S., King, E., Cortina, J. (Eds.), *Big data at work: The data science revolution and organizational psychology* (pp. 273–309). Routledge Academic.

Langer, M., Mast, M. S., Meyer, B., Maass, W., & König, C. J. (2019). Research in the era of sensing technologies and wearables. In R. Landers (Ed*.), The Cambridge handbook of technology and employee behavior* (pp. 806–835). Cambridge University Press.

Luciano, M. M., Mathieu, J. E., Park, S., & Tannenbaum, S. I. (2018). A fitting approach to construct and measurement alignment: The role of big data in advancing dynamic theories. *Organizational Research Methods*, *21*(3), 592-632. https://doi.org/10.1177/1094428117728372

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*(3), 356–376. https://doi.org/10.5465/amr.2001.4845785

Matusik, J. G., Heidl, R., Hollenbeck, J. R., Yu, A., Lee, H. W., & Howe, M. (2019). Wearable bluetooth sensors for capturing relational variables and temporal variability in relationships: A construct validation study. *Journal of Applied Psychology*, *104*(3), 357-387. https://doi.org/10.1037/apl0000334

Milosevic, M., Jovanov, E., & Frith, K. H. (2013). Research methodology for real-time stress assessment of nurses. *Computers Informatics Nursing*, *31*(12), 615-621. https://doi.org/10.1097/CIN.0000000000000011

Mønster, D., Håkonsson, D. D., Eskildsen, J. K., & Wallot, S. (2016). Physiological evidence of interpersonal dynamics in a cooperative production task. *Physiology and Behavior*, *156*, 24–34. https://doi.org/10.1016/j.physbeh.2016.01.004

Nickel, P., & Nachreiner, F. (2003). Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload. *Human Factors*, *45*(4), 575-590. https://doi.org/10.1518/hfes.45.4.575.27094

Roma, P. G., Hursh, S. R., & Hudja, S. (2016). Hypothetical purchase task questionnaires for behavioral economic assessments of value and motivation. *Managerial and Decision Economics*, *37*(4-5), 306-323. https://doi.org/10.1002/mde.2718

Roma, P. G., Reed, D. D., DiGennaro Reed, F. D., & Hursh, S. R. (2017). Progress of and prospects for hypothetical purchase task questionnaires in consumer behavior analysis and public policy. *The Behavior Analyst*, *40*(2), 329-342. https://doi.org/10.1007/s40614-017-0100-2

Salas, E., Grossman, R., Hughes, A. M., & Coultas, C. W. (2015). Measuring team cohesion: Observations from the science. *Human Factors*, *57*(3), 365-374. https://doi.org/10.1177/0018720815578267

Strang, A. J., Funke, G. J., Russell, S. M., Dukes, A. W., & Middendorf, M. S. (2014). Physio-behavioral coupling in a cooperative team task: Contributors and relations. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 145– 158. doi:10.1037/a0033125

Thayer, J. F., Åhs, F., Fredrikson, M., Sollers III, J. J., & Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, *36*(2), 747-756. https://doi.org/10.1016/j.neubiorev.2011.11.009

Thayer, J. F., Hansen, A. L., Saus-Rose, E., & Johnsen, B. H. (2009). Heart rate variability, prefrontal neural function, and cognitive performance: the neurovisceral integration perspective on self-regulation, adaptation, and health. *Annals of Behavioral Medicine*, *37*(2), 141-153. https://doi.org/10.1007/s12160-009-9101-z

Appendix A

Surveys for MSTC Instructors

TEAM PERFORMANCE

Based on your observations of the participating squad today (i.e., Soldiers wearing the Empatica wristbands), please answer the following questions using the scale shown below.

○	○	○	○	○
Not at All		**To a Moderate Extent**		**To a Very Great Extent**

1.	To what extent did the squad accomplish its primary goals today?	○ ○ ○ ○ ○
2.	To what extent were the important tasks for today done in a high quality fashion?	○ ○ ○ ○ ○
3.	To what extent did the squad perform well today?	○ ○ ○ ○ ○
4.	To what extent were the important tasks for today done in a timely fashion?	○ ○ ○ ○ ○

COHESION

Based on your observations of the participating squad today (i.e., Soldiers wearing the Empatica wristbands), please rate the following statements using the scale shown below.

○	○	○	○	○
Strongly Disagree	**Disagree**	**Neither Agree nor Disagree**	**Agree**	**Strongly Agree**

1.	The squad was unified in its task focus.	○ ○ ○ ○ ○
2.	The squad cared about each other.	○ ○ ○ ○ ○
3.	The squad had a shared sense of task importance.	○ ○ ○ ○ ○
4.	The squad members had good relationships with each other.	○ ○ ○ ○ ○
5.	The squad was committed to its task(s).	○ ○ ○ ○ ○
6.	The squad members enjoyed each other's company.	○ ○ ○ ○ ○
7.	The squad has a shared sense of purpose and commitment.	○ ○ ○ ○ ○

Appendix B

Surveys for MSTC Students

TEAM PERFORMANCE

Answer the below questions with regard to what happened today with your squad.

○	○	○	○	○
Not at All		**To a Moderate Extent**		**To a Very Great Extent**

1.	To what extent did your squad accomplish its primary goals today?	○ ○ ○ ○ ○
2.	To what extent were the important tasks for today done in a high-quality fashion?	○ ○ ○ ○ ○
3.	Taking everything into consideration, to what extent did your squad perform well today?	○ ○ ○ ○ ○
4.	To what extent were the important tasks for today done in a timely fashion?	○ ○ ○ ○ ○

COHESION

Based on your experience working with your squad today, please rate the following statements using the scale shown below.

○	○	○	○	○
Strongly Disagree	**Disagree**	**Neither Agree nor Disagree**	**Agree**	**Strongly Agree**

1.	Our squad was unified in its task focus.	○ ○ ○ ○ ○
2.	Our squad cared about each other.	○ ○ ○ ○ ○
3.	Our squad had a shared sense of task importance.	○ ○ ○ ○ ○
4.	Our squad members had good relationships with each other.	○ ○ ○ ○ ○
5.	Our squad was committed to its task(s).	○ ○ ○ ○ ○
6.	Our squad members enjoyed each other's company.	○ ○ ○ ○ ○
7.	Our squad has a shared sense of purpose and commitment.	○ ○ ○ ○ ○

Appendix C

MSTC Survey Results

	Day 1			Day 2			Day 3			Day 4		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Team Performance – Instructor Rated	—	—	—	—	—	—	3	4.33	0.65	6	4.38	0.71
Team Performance – Student Rated	6	4.58	0.50	6	4.63	0.49	6	4.75	0.44	6	4.75	0.44
Team Cohesion – Instructor Rated	—	—	—	—	—	—	3	4.86	0.36	6	4.46	0.71
Team Cohesion – Student Rated	6	4.52	0.51	6	4.62	0.49	6	4.71	0.46	6	4.71	0.46

Note. All responses were on a five-point agreement scale (1 = strongly disagree, 5 = strongly agree).