A Multimodal Social Signal Processing Approach to Team Interactions

Lehmann-Willenbrock, Nale; Hung, Hayley

DOI
10.1177/10944281231202741

Publication date
2024

Document Version
Final published version

Published in
Organizational Research Methods

Citation (APA)
Lehmann-Willenbrock, N., & Hung, H. (2024). A Multimodal Social Signal Processing Approach to Team Interactions. *Organizational Research Methods*, *27*(3), 477-515. https://doi.org/10.1177/10944281231202741

Article

Organizational Research Methods
2024, Vol. 27(3) 477–515
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10944281231202741
journals.sagepub.com/home/orm
Sage

A Multimodal Social Signal Processing Approach to Team Interactions

Nale Lehmann-Willenbrock[1]
and Hayley Hung[2]

Abstract

Social signal processing develops automated approaches to detect, analyze, and synthesize social signals in human–human as well as human–machine interactions by means of machine learning and sensor data processing. Most works analyze individual or dyadic behavior, while the analysis of group or team interactions remains limited. We present a case study of an interdisciplinary work process for social signal processing that can develop automatized measures of complex team interaction dynamics, using team task and social cohesion as an example. In a field sample of 25 real project team meetings, we obtained sensor data from cameras, microphones, and a smart ID badge measuring acceleration. We demonstrate how fine-grained behavioral expressions of task and social cohesion in team meetings can be extracted and processed from sensor data by capturing dyadic coordination patterns that are then aggregated to the team level. The extracted patterns act as proxies for behavioral synchrony and mimicry of speech and body behavior which map onto verbal expressions of task and social cohesion in the observed team meetings. We reflect on opportunities for future interdisciplinary or collaboration that can move beyond a simple producer–consumer model.

Keywords

machine learning and AI, types of research design, field research, types of research design, time series, longitudinal and related approaches

Many research questions in organizational research relate to how members of organizations behave during dynamic social interactions (LeBaron et al., 2018). Among social interaction phenomena in organizations, team interactions are particularly complex and often puzzling for organizational researchers (Waller & Kaplan, 2018). However, researchers need to understand behavioral team interactions in order to understand emergent team phenomena such as intra-team trust, collaborative

[1] University of Hamburg, Hamburg, Germany
[2] Delft University of Technology, Delft, The Netherlands

Corresponding author:
Nale Lehmann-Willenbrock, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany.
Email: nale.lehmann-willenbrock@uni-hamburg.de

sensemaking, or team cohesion, all of which originate in team behavioral dynamics. Several scholars have pointed to the benefits of recording audiovisual team interaction data in order to capture rich data on team dynamics over time (e.g., Kozlowski, 2015; Waller & Kaplan, 2018). Empirical studies adopting audiovisual recording and analysis have made important contributions to understanding the micro-level behavioral dynamics that characterize successful team interactions (e.g., Hoogeboom & Wilderom, 2020; Lehmann-Willenbrock & Chiu, 2018; Uitdewilligen et al., 2018). Yet, behavioral team phenomena continue to be difficult to capture. If researchers go to the trouble of gathering audiovisual data on actual team behavior rather than relying on survey-based proxies of behavior (for a detailed critique, see Lehmann-Willenbrock & Allen, 2018), they need substantial resources for annotating the data and quantifying behavioral patterns. Moreover, team behavioral dynamics are fluid and often change from one minute to the next, which requires "high-resolution" research methods with high sampling rates (Klonek et al., 2019).

This paper presents a social signal processing approach that can automatically detect "high-resolution" behavioral team processes from sensor data that is combined with a machine learning algorithm. This approach moves beyond the state of the art in team interaction analysis in several ways. First, social signal processing can incorporate multiple social signals that occur simultaneously during team interactions, whereas researchers pursuing quantitative team interaction analysis typically focus on only one modality and study sequential verbal or nonverbal interaction behavior (for an overview, see Keyton, 2018). Second, "high resolution" is a debatable term in the literature. The state of the art often considers verbal utterances as the smallest temporal unit (cf., Klonek et al., 2019), but behavioral phenomena in complex social collectives such as teams also occur at a much more fine-grained, sub-utterance level of nonverbal behavior (see Müller et al., 2019). Third, the state of the art in team interaction analysis requires intense human effort, in terms of the many hours that go into annotating the interaction data. Social signal processing develops algorithms that can adequately predict such human annotations (or ground truth labels), with the aim of automating these predictions.

With the growth in artificial intelligence in the last decade, many human behaviors can be measured reliably by state-of-the-art machine learning techniques (see Rudovic et al., 2017). The main premise of machine learning in its most basic form is to learn a mapping from some data to an expected outcome or "label." Mathematically speaking, this involves minimizing an error between a prediction and an actual outcome by adjusting the parameters of this mapping function. When the learning task involves predicting some aspect of a human's social behavior given some behavioral input data, this is known as the research domain of social signal processing, defined as "the computing domain aimed at modeling, analysis, and synthesis of social signals in human–human and human–machine interactions" (Vinciarelli, 2017).

Social signals are constructs that are generated from a constellation of measurable behavioral cues displayed during social interactions, such as facial expressions, gaze, body posture, movement, gestures, and vocal expressions (e.g., speech rate) that produce a response in others (e.g., team members). The data used in social signal processing are often multimodal social signals that have been captured by sensors in the local environment, capturing, for example, (1) video from cameras, (2) audio from microphones, or (3) bodily movement or physiological data from wearable sensors. Relatively simple social cues such as facial action units from video or sentiment from automatically transcribed text can be extracted without additional human annotation, often reaching acceptable degrees of reliability (for an overview, see Burgoon et al., 2017). To study more complex, multimodal social behaviors such as team processes, machine learning algorithms require training by humans to detect meaningful behaviors. Hence, the final ingredient for training a machine to interpret more complex group or team constructs is human intervention. This may initially trigger some disappointment for researchers who are looking for off-the-shelf technology to capture their constructs of interest. However, once there are sufficient interdisciplinary collaborations to generate more robust machine learning approaches, this will catalyze our understanding of team dynamics and temporal linkages between different team processes (and other behavioral interaction phenomena, including leadership, e.g., Fischer et al., 2020; Hemshorn de Sanchez et al., 2022).

In this paper, we showcase the potential of an interdisciplinary social signal processing approach for obtaining new insights into systematic behavioral patterns in teams and other interacting social collectives in organizations. We apply this approach in a field sample of project team meetings, during which we recorded numerous social signals (audio, video, and movement) and annotated the teams' verbal interaction with the aim to automatically predict moments of high or low cohesion in the meeting from team patterns of extracted social signals. We specify the requirements for multimodal social signal data gathering, explain how to select appropriate time windows, calculate measures of behavioral mimicry based on multimodal sensor data at the team level, and investigate to what extent automatically extracted behavioral mimicry can predict cohesive team interaction behaviors. We discuss the potential as well as the shortcomings and substantial need for additional interdisciplinary or even transdisciplinary work on automatic behavioral modeling approaches to dynamic team interaction phenomena, in the hope that this paper will inspire others to pool their expertise and embrace interdisciplinary research collaboration opportunities in this area.

Detecting Behavioral Mimicry and Predicting Cohesive Team Interaction Using Social Signal Processing

At the outset, we need to clarify that machine learning, which is the method used in social signal processing for making sense of social signals gathered from various sensors, is different from traditional regression analytical methods in several key ways. Statistical modeling approaches such as regression center on the assumption that data are generated by a given stochastic data model. In contrast, supervised machine learning typically uses algorithms for modeling with the assumption that the process that generated the data given a particular ground truth label is unknown. This distinction between machine learning and statistical modeling has been discussed in depth by Breiman (2001), along with the limitations of statistical modeling approaches. However, machine learning algorithms can and do also utilize regression and logistic regression techniques as a basic tool for prediction tasks. For example, machine learning uses linear or logistic regression analysis for supervised learning, in order to fit a function on the available data (e.g., James et al., 2021). One reason for using supervised learning methods such as logistic regression, which we also use in the current study, is that it somewhat circumvents the challenge of interpretability of machine learning models (i.e., explainable machine learning; e.g., Arrieta et al., 2020). Of note, next to supervised learning approaches such as the current study, machine learning can also employ unsupervised, data-driven approaches to discover patterns in the data as well as meta-learning (e.g., Hospedales et al., 2022; Vanschoren, 2019).

Another way to look at the difference between machine learning and statistical modeling approaches such as regression is to consider the types of analysis that can be achieved by each approach. Regression analysis yields insights into generalized behavior that may be more or less helpful for productive collaboration in teams. In comparison, machine learning enables a more personalized analysis approach. Notably, sometimes machine learning produces different findings than regression analysis. This is because machine learning models are capable of accounting for much higher complexity. Moreover, an important distinction between regression analysis and machine learning is that machine learning intends to automate analysis, whereas regression analysis does not. In other words, the promise of machine learning methods in the area of social signal processing is that we will eventually be able to automatically detect meaningful social interaction phenomena, including complex team constructs such as cohesion.

Social signal processing approaches quickly yield a wealth of data. However, in order to advance organizational research using such methods, the extraction and interpretation of social signals should be guided by theoretical constructs. In our case, we focus on team cohesion, which has been researched more extensively than any other team phenomenon. Several meta-analyses support linkages between cohesion and team performance (e.g., Beal et al., 2003; Castaño et al., 2013; Chiocchio & Essiembre, 2009; Evans & Dion, 1991). Cohesion can be defined as a shared bond or attraction among team members that holds the team together and is grounded in task-based or social aspects of team membership (Casey-Campbell & Martens, 2009). Task cohesion refers to a general orientation toward achieving the group's goals and objectives, whereas social cohesion refers to a general orientation toward developing and maintaining social relationships within the group (e.g., Carron et al., 1985).

As is the case for most team constructs, prior empirical work on team cohesion has predominantly relied on self-report survey measures, often with a cross-sectional approach that is difficult to reconcile with the conceptual understanding of cohesion as a dynamic group property (e.g., Kozlowski & Chao, 2012; Salas et al., 2015). In response, recent work has embraced a longitudinal approach to the study of cohesion in teams (Acton et al., 2020; Hill et al., 2019). However, while contributing interesting insights into the development of team members' self-reported perceptions of cohesion over time, these studies still offer little in the way of actual behavioral expressions of cohesion. To address this shortcoming in the extant literature, scholars have pointed to audio/video recording of team interactions in order to capture rich data on team dynamics over time (Kozlowski, 2015; Kozlowski & Chao, 2012, 2018; Santoro et al., 2015).

We need to emphasize here that the issue of mismatching theory (which specifies behavioral constructs) and empirical investigations of team constructs (which continue to rely on survey-based behavioral proxies) is a pervasive problem in the literature on team processes, as discussed in detail elsewhere (e.g., Kozlowski & Chao, 2018; Klonek et al., 2019; Lehmann-Willenbrock & Allen, 2018). Note that similar discussions are currently led in the leadership literature (Banks et al., in press; Hemshorn de Sanchez et al., 2022). We ask readers to keep an open mind about the potential of social signal processing approaches to address this core issue in the broader organizational behavior literature, and we use the example of team cohesion for demonstration purposes in this regard.

To clarify our methodological focus within the team process literature, we further need to emphasize the distinction between team emergent states (i.e., the traditional conceptual approach to cohesion) and the behavioral processes of emergence that lead to these states. Insights into the latter have much more explanatory value for understanding the core behavioral mechanisms of dynamic team constructs, for example, when aiming to understand how cohesion develops and changes over time (cf., Kozlowski & Chao, 2018). Pentland and Heibeck (2008) discussed the promise of sensor data to understand social phenomena. Since then, sensor data have mainly been used to predict static measures of team constructs (e.g., predicting self-report surveys of cohesion), even in computer science. Moreover, there is an overreliance on controlled, "clean" laboratory settings in social signal processing, given that high-performing machine learning algorithms are easier to achieve in these settings (for an overview, see Müller et al., 2019). Instead, we propose an application to real-life, messy, temporally dynamic processes in teams (cf., Klonek et al., 2019) and increase the challenge from the computer science perspective as well by aiming to predict much more fine-grained instantiations of team cohesion expressions within dynamic team interactions. Müller et al. (2019) provide a detailed discussion of the incompatibility between highly granular sensor data on the one hand and the static, a-temporal nature of traditional quantitative methods on the other hand (as in the case of the survey-based literature on team cohesion). To address this issue, we need to match the criterion for training our machine learning algorithm much more closely to the high resolution of social signals exchanged during team interactions.

Providing a Ground Truth for Behavioral Expressions of Team Cohesion

A criterion is needed for training and developing algorithms that automatically detect and combine social signals and for evaluating their performance. This criterion is called the ground truth for the construct of interest. It requires labeled data that can be taken as definitive and against which the automated system can be measured and trained (e.g., Pantic et al., 2011). To establish such a ground truth, usually behavioral data are collected for which a human expert has provided a judgment of what the machine should predict. This judgment is the “reference” or “ground truth.” A ground truth can be provided by annotating (or rating or coding) observed behaviors during a social interaction, or it can come from self-reported or other-reported survey measures, external ratings, or outcomes such as team performance. Machine learning algorithms try to minimize the discrepancy between the predicted label and the reference or ground truth. Given sufficient training data, such an algorithm can then be used to automatically analyze new data that is similarly distributed. In our case, this means that given acceptable performance, an algorithm developed to detect cohesion can eventually be applied to new datasets for automatically detecting team cohesion without requiring human annotation effort.

Providing a ground truth at the behavioral event level can be a challenge of its own. In our case, we are looking for behavioral expressions of team cohesion as a starting point for a social signal processing approach. Of note, our aim is to pinpoint indicators of cohesion in the team interaction process, whereas most of the extant literature has investigated perceptions of emergent cohesion (i.e., as the result of prolonged team interactions, although the latter are typically not investigated). Hence, we take some liberties in extrapolating from the extant literature. Some of the extant survey-based research points to observable behavioral indicators of task and social cohesion, respectively. Table 1 provides an overview of these findings and shows how we operationalized each finding at the behavioral event level in team interactions, drawing from the literature on behavioral interactions in team meetings.

For example, Zaccaro and colleagues (1995) discussed how task cohesive teams devote more time to action planning. In terms of observable verbal behaviors, action planning expresses concrete intentions to act, such as “I’ll take care of this today” or “We’ll have this done by Friday” (e.g., Kauffeld et al., 2018). As further detailed in Table 1, we also include statements that express taking responsibility, such as “This is our job as a team” (e.g., Kauffeld & Lehmann-Willenbrock, 2012), or that signal positivity, such as “This could really work” (Lehmann-Willenbrock et al., 2017). In line with the literature (Table 1), we also consider procedural statements (e.g., Lehmann-Willenbrock et al., 2013) such as distributing tasks, prioritizing, and goal orientation statements as expressions of task cohesion. Regarding social cohesion, we extrapolate from the extant literature as highlighted in Table 1 and consider relational communication such as supportive statements, praise, active listening, expressions of feelings (e.g., Kauffeld & Lehmann-Willenbrock, 2012), humor and laughter (Lehmann-Willenbrock & Allen, 2014), and statements encouraging other members to participate (e.g., “Lisa, what do you think?”) as behavioral indicators of social cohesion in team interactions.

Mimicry and Synchrony of Social Signals

The social signal processing community has frequently considered the role of mimicry or synchrony in social interactions. Mimicry, the interdependence of interacting partners’ behaviors, occurs in almost any social interaction when different individuals converge in their social signals (e.g., Delaherche et al., 2012; Duffy & Chartrand, 2015). Of note, group interactions are considerably more dynamic and complex than dyadic interaction scenarios (e.g., Lehmann-Willenbrock et al., 2017). In thin slices of behavior within the team interaction stream, we show how automatically

Table 1. Expression of Task and Social Cohesion in Team Interactions.

Task Cohesion		Behavioral Expression	Examples
Zaccaro et al. (1995)	Highly task cohesive teams devote more time to action planning	Action planning	"I'll go see the project lead on Monday and ask him."
Eys and Carron (2001)	Clarity of responsibilities positively related to perceived task cohesion	Taking responsibility	"The initiatives have to come from us."
Carron et al. (1985); Whitton and Fletcher (2014)	Shared responsibility as an expression of task cohesion		
Vanhove and Herian (2015)	Optimism as a potential indicator of task cohesion	Expressing positivity	"That sounds like a great idea!"
West et al. (2009)	Team positivity positively linked to team cohesion		
Brawley et al. (1987)	Higher task cohesion when group task requires organizing task requirements	Task distribution	"Matthew, please make a note of that."
Shields et al. (1995)	Coherent, shared set of team priorities may facilitate task cohesion	Prioritizing	"We should do that first."
Zaccaro et al. (1995)	Team members need to prioritize tasks during planning in order to achieve cohesion		
Salas et al. (2015)	Individual goal orientation likely to influence team task cohesion	Goal orientation	"Let's get back to our topic, which was…"
Carron et al. (1985); Whitton and Fletcher (2014)	Task cohesion as general orientation toward group goal achievement		
Acton et al. (2020)	Team-level goal orientation predicts task cohesion		
Social Cohesion		**Behavioral Expression**	**Examples**
Christensen et al. (2006)	Attitude of mutual support as a characteristic of (socially) cohesive groups	Providing support	"I fully agree."
Braaten (1990)	Cohesive groups are characterized by a social climate of support and caring		
Burlingame et al. (2001)	Listening and empathy are characteristic for highly cohesive (therapy) groups	Active listening	Strong nodding: "Mmh, yes."

(continued)

Table 1. (continued)

Social Cohesion		Behavioral Expression	Examples
Carmody et al. (2017)	Active listening may facilitate team rapport and cohesion		
Lott and Lott (1965)	Cohesion in terms of positive judgments that team members make of one another	Offering praise	"You did such a great job earlier…"
Hüffmeier and Hertel (2011)	Mutual provision of social recognition should foster social cohesion in teams		
Greatbatch and Clark (2003); Holmes and Marra (2006); Scogin and Pollio (1980); Terrion and Ashforth (2002)	Shared humor can promote social cohesion Leaders' humor use builds social cohesion Laughter as a display of social cohesion	Humor; laughter	
Burlingame et al. (2001)	Emotional expression as a cue for relational cohesion in (therapy) groups	Expressing feelings	"I don't feel comfortable with the idea that…"
Glass and Benshoff (2002)	Mutual encouragement and inclusion as an expression of social cohesion	Encouraging participation	"You haven't said anything to that yet, Thomas. Why don't you say something."
Hüffmeier and Hertel (2011)	Encouragement should foster social cohesion		

detected behavioral mimicry based on social signal data can be captured at the group level to predict behavioral expressions of task and social cohesion in team interactions.

Humans have a natural tendency to mimic one another's facial expressions, emotions, speech characteristics and patterns, and motor movements such as posture and gestures (for an overview, see Chartrand & Van Baaren, 2009). Different types of mimicry can be considered behavioral (Chartrand & Lakin, 2013). In other words, mimicry can occur across a range of different behavioral modalities such as gestures, postures, facial expressions, or vocal expressions. In teams, behavioral mimicry occurs when two or more members exhibit similar behavioral patterns within a short time window (i.e., in quick temporal succession, such as a few seconds). In our application example, we use social signal processing to model multimodal mimicry among team members as an underlying mechanism that may help us automatically detect cohesive team interaction behaviors in project team meetings. In other words, when team members mimic their nonverbal expressions and move in sync, this can point to cohesive team interactions.

Research on dyadic interactions has identified several social benefits of mimicry, including increased empathy, bonding, and positive affect (e.g., Stel & Vonk, 2010; Tschacher et al., 2014). Previous research has also found connections between cohesion and behavioral synchrony, defined as the spontaneous rhythmic and temporal coordination of actions between two or more participants (Chartrand & Lakin, 2013; Delaherche et al., 2012; Lakin, 2013; Mayo & Gordon, 2020). Hoehl and colleagues (2021) review and discuss the manifold benefits of synchrony in human interactions, including bonding. Wilson and Gos (2019) found that dyads moving in synchrony are perceived as more cohesive. In the context of large social groups, Jackson and colleagues (2018) showed that behavioral synchrony in movement and arousal increased cohesion. These findings suggest that a social signal processing approach to team interactions should capture mimicry or synchrony.

Whereas some social signal processing work has focused on a single modality only when modeling behavioral mimicry (Nanninga et al., 2017), additional behavioral modalities may need to be considered. Verbal indicators of task and social cohesion (Table 1) may be accompanied by a wealth of nonverbal signals spanning different modalities such as body movement and paralinguistic features (e.g., voice pitch). Taking a step toward multimodal integration, one previous study observed that dyadic mimicry of movement (i.e., accelerometer data) and speech signals, aggregated to the group level, was positively correlated with self-reported task cohesion at the day level (Zhang et al., 2018). Of note, this previous research captured mimicry by aggregating the similarity of participants' social signals across 10-min periods, which is likely an overly rough time resolution that cannot account for more swift developments and changes in team interaction behaviors. Next, we investigate how social signal mimicry in different behavioral modalities and within fine-grained temporal windows maps onto expressions of task and social cohesion observed in a sample of project team meetings in the field.

Data Gathering

Data were gathered during regular project team meetings at a large software organization in the Netherlands. Our study was endorsed by the local ethics committee as well as the organization's legal department. Participation was voluntary and subject to informed consent by all team members. Data confidentiality was guaranteed. Participants retained the right to withdraw from the data gathering at any time and to have their data deleted upon request. Twenty-five regular meetings were recorded. Meeting size ranged from 3 to 8, with an average of 4.6 attendees. Participants were 64% male, 37 years old on average, with an organizational tenure of 8.6 years and an average team tenure of 2.5 years. To enable multimodal behavioral data gathering, we equipped an on-site meeting room (see Figure 1) with multiple cameras, headset microphones, a microphone array

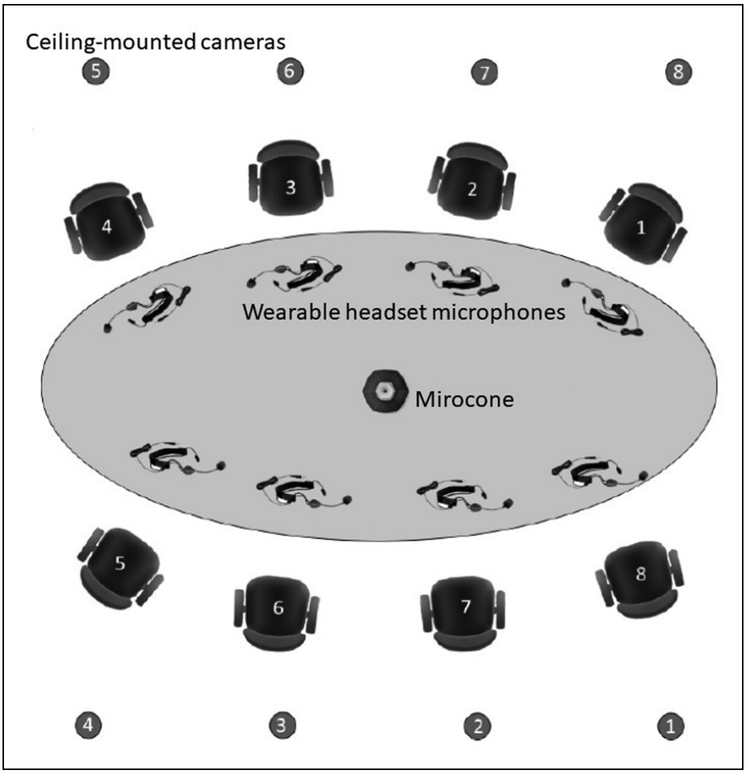

Figure 1. Schematic overview of the on-site meeting room setup showing camera locations, with each camera capturing an individual meeting attendee seated across each camera (1 through 8); wearable headset microphones, one per individual meeting attendee; and the Microcone (microphone array) placed in the center of the boardroom table. Figure adapted from Nanninga et al., (2017). In addition, each attendee wore a custom-built sensor badge around the next to track their motion during the meeting.

(Microcone; Mast et al., 2015), and a wearable sensor badge. Eight cameras (ABUS HDCC72510, which record at a resolution of 1,920 × 1,080 and at 25 fps), each centered on the seat on the opposite side of the room, were mounted overhead in order to avoid distracting the participants. The Microcone was placed in the center of the table. All individual meeting attendees were outfitted with a microphone (Sennheiser wireless headsets, version ew 152 G3) to record their speech and a custom-built sensor badge worn around the neck measuring torso acceleration at three orthogonal directions of motion. Out of the 25 recorded meetings, 14 had complete data available across all modalities (Table 2).

Developing Ground Truth Labels

As a ground truth and criterion for training and evaluating our machine learning algorithm, we identified moments of high or low task/social cohesion across 2-min segments of each meeting (in line with previous work on convergent group phenomena; Barsade, 2002). These labels of high or low task and social cohesion were based on the annotated verbal behavior that occurred during each 2-min segment (see Figure 2). Two intensively trained coders annotated the entire stream of verbal interactions using the act4teams coding scheme for team interactions (e.g., Kauffeld et al., 2018), yielding a total of 19,979 verbal utterances. An utterance or sense unit is the smallest

Table 2. Available Sensor Data for Each Modality and Analyzed Meeting Segments Across All Observed Team Meetings.

Meeting	Total Number of Participants	Annotated Verbal Interaction	Usable Video Data	Participants with Accelerometer Data	Meeting Duration (Hours:Minutes: Seconds)	Number of Analyzed Meeting Segments
1	5			N	01:13:35	37
2	4			N	01:24:41	42
3	5	N				
4	8			7	00:59:17	31
5	6			4	00:50:57	26
6	4			3	01:28:51	44
7	3			3	00:39:47	20
8	3			2	00:22:12	11
9	5			3	00:24:13	12
10	5			N	00:55:28	27
11	6			6	00:46:48	23
12	5			5	00:42:25	21
13	4			3	00:55:38	27
14	4			N	00:36:33	18
15	5			4	00:52:56	27
16	5			5	00:28:28	14
17	3	N				
18	6			4	00:37:35	18
19	3		N	3	01:12:22	36
20	6			4	00:42:23	21
21	6			6	00:49:03	24
22	5		N	5	00:48:30	24
23	4		N	3	00:59:52	30
24	3		N	3	01:11:11	36
25	3			3	01:02:28	31

Note: N: data was not available, not time-synchronized with the other modalities, or participants walked outside the camera angle. Meeting 8 was excluded concerning the accelerometer data as it only included accelerometer data for two participants, which does not constitute a group (Moreland, 2010).

speech segment that expresses a complete thought (Bales, 1950). This is often the same as a single sentence, but it can also be a single word (e.g., “Uh-huh” for active listening), leading to a very fine-grained analysis. The final column in Table 1 shows examples of verbal utterances. Like many contemporary studies of behavioral team interactions, the coders used software (specifically, INTERACT software) to annotate directly from the video without needing to transcribe. This creates time stamps with onset and offset times as well as duration for each annotated behavior (for an overview, see Lehmann-Willenbrock & Allen, 2018). Nine randomly selected meetings were annotated twice in order to establish inter-rater reliability for the utterance annotations that were the basis for our ground truth labels ($\kappa = .80$).

To get from the annotated verbal utterances to labels that can be used for social signal processing, we pursued a typical social signal processing approach and labeled the team interaction data using binary ground truth labels for high versus low task and social cohesion, respectively. For each 2-min meeting segment, we computed a proportion of task cohesion and a proportion of social cohesion, based on the verbal behaviors that occurred during each meeting segment (as

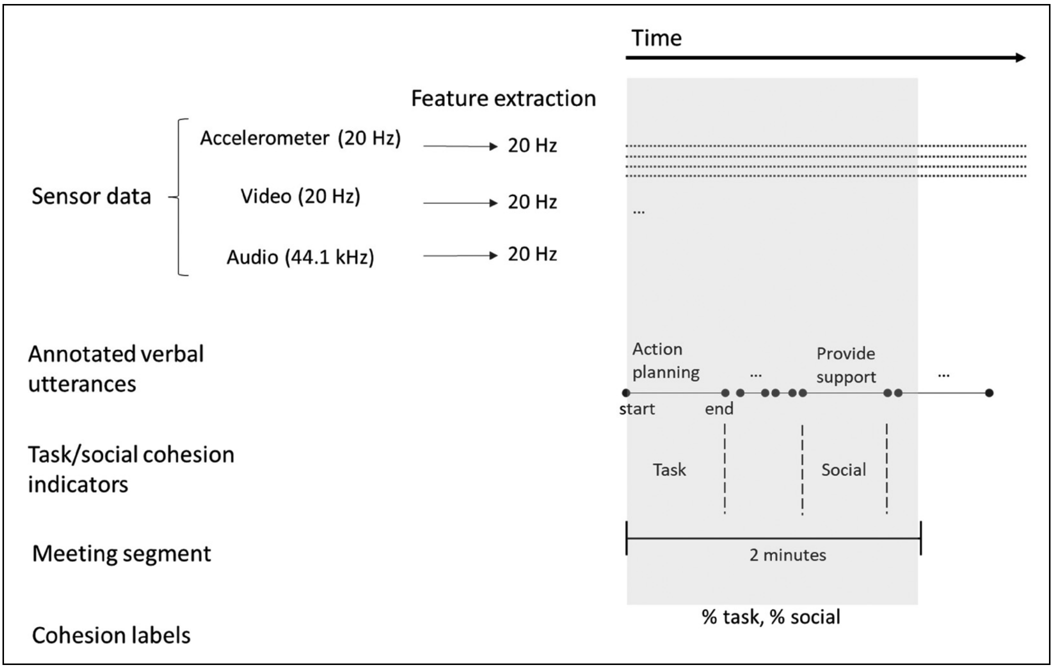

Figure 2. General data gathering framework. Type of data shown on the left and temporal resolution shown toward the right. Sensor data including sampling rates. All features were extracted for 2-min meeting segments within each team meeting (as illustrated by the highlighted section). We extracted 720 features in total (specifics regarding feature extraction are detailed next in the paper). For each 2-min segment, we obtained the ground truth (i.e., cohesion labels) from the annotated verbal utterances, as illustrated in the lower part of the figure. Cohesion labels were obtained at the meeting segment level from the percentages of task and social cohesion behaviors within the respective meeting segment.

illustrated in Figure 2). To obtain these labels, we focused on the annotated utterances that indicate task and social cohesion as described in Table 1, third column. From these annotations, for each 2-min segment in each team meeting, we computed the proportion of time that the team spent on these behaviors. We converted these proportions into labels for high and low task and social cohesion by dividing the duration of the verbal behaviors that express task and social cohesion by the duration of all verbal behaviors annotated within each 2-min segment. This produced social and task cohesion values for all 2-min segments of each team meeting which ranged between 0 and 1. Among these values, we defined a “high” label as a value in the top 25% of the distribution and a “low” label as a value in the bottom 25% of the distribution, respectively (for a similar approach, see Hung & Gatica-Perez, 2010). Of note, high task cohesion labels occurred sparsely, whereas moments labeled as high social cohesion were more frequent (see Appendix A for more detail).

Data Processing

In the domain of social signal processing and in machine learning more generally, major time and effort are devoted to pre-processing the gathered sensor data and extracting relevant features from this data (e.g., Khalid et al., 2014; Ramírez-Gallego et al., 2017; Vinciarelli et al., 2009). Hence, we discuss these two aspects of developing a machine learning algorithm for team behavior analysis in detail. Figure 3 provides an overview of our workflow to extract relevant features and

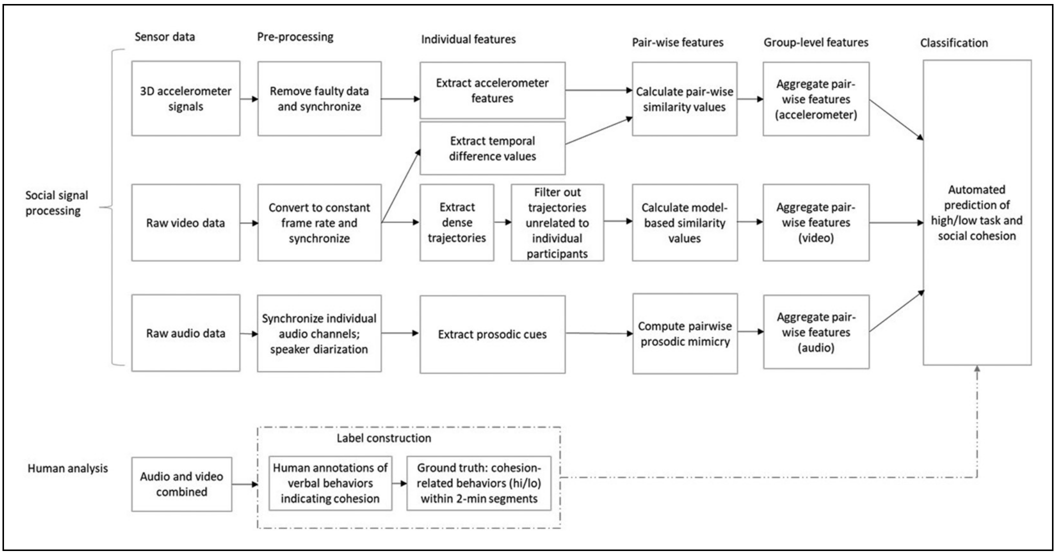
Sensor data
Pre-processing
Individual features
Pair-wise features
Group-level features
Classification
Social signal processing
3D accelerometer signals
Remove faulty data and synchronize
Extract accelerometer features
Extract temporal difference values
Calculate pair-wise similarity values
Aggregate pair-wise features (accelerometer)
Raw video data
Convert to constant frame rate and synchronize
Extract dense trajectories
Filter out trajectories unrelated to individual participants
Calculate model-based similarity values
Aggregate pair-wise features (video)
Automated prediction of high/low task and social cohesion
Raw audio data
Synchronize individual audio channels; speaker diarization
Extract prosodic cues
Compute pairwise prosodic mimicry
Aggregate pair-wise features (audio)
Human analysis
Label construction
Audio and video combined
Human annotations of verbal behaviors indicating cohesion
Ground truth: cohesion-related behaviors (hi/lo) within 2-min segments

Figure 3. Flow chart of the work process for automatically detecting cohesion in meetings using a social signal processing approach.

automatically predict cohesion labels. The top part of Figure 3 refers to the machine learning approach to the observed team meetings that can account for multiple, simultaneous modalities in team behavior (movement, video, and audio). From the three different modalities, we calculated group-level similarity (i.e., mimicry) values among team members. These values were used to predict high/low cohesion labels. The bottom part of the figure refers to the human analysis of these team meetings. The dotted lines illustrate how we obtained ground truth labels for task and social cohesion from the human annotations of the observed team meetings. These labels served as criteria for training and evaluating the machine learning algorithm for automatically detecting cohesion.

Figure 3 also indicates key differences between social signal processing and the state of the art in quantitative team interaction analysis. The latter is sketched at the bottom of Figure 3 and essentially requires two steps: (1) gather behavioral team data and (2) spot behaviors of interest using trained human annotators (for more detail, see Keyton, 2018; Lehmann-Willenbrock & Allen, 2018). This means that the temporal windows for analyzing the team interaction flow are much more coarse than in social signal processing, where the behavioral units are milliseconds or picture frames rather than seconds or minutes. Furthermore, the bottom path illustrated in Figure 3 falls short of analyzing the rich multimodal quality of behavioral interaction dynamics in teams, due to the extensive human annotation effort that would be required to address this. Finally, whereas the ultimate goal in social signal processing is to automatically obtain reliable predictions of meaningful behavior from sensor data, this automation is not possible in quantitative interaction analysis (see also Allen et al., 2017, for a detailed discussion of differences in workflows when social scientists vs. computer scientists are analyzing team interactions).

We pre-processed the sensor data from each modality per team member into time series data across 2-min segments within each meeting. For each modality, we extracted features and captured behavioral synchrony and convergence across these 2-min segments (i.e., the same timeframe that was used to establish ground truth labels; see Figure 2). For the code for our feature extraction and combination, we used a combination of Python and Matlab. While we are not permitted to openly share the code, extracted features, and the underlying raw behavioral measures given privacy concerns and legal restrictions at the organization where we gathered our field data, interested readers can contact us for more information on an individual basis.

Extracting Multimodal Features

We extracted time series data from (1) the audio data (voice intensity, fundamental frequency, speech rate, and first 12 mel frequency cepstral coefficients [MFCCs]); (2) the video data (thresholded pixel-wise difference between image frames and the Histogram of Oriented Gradient [HOG] part of the dense trajectories feature vector); and (3) the accelerometer data (absolute x, y, z, and magnitude, from which six spectral and two statistical representations, namely, mean and variance, were generated—thus, $4 \times (6 + 2)$ features).

Extracting Vocal Features from the Audio Data

From the Microcone (see Figure 1) audio data, we obtained audio segments separated by speaker. Using the open-source program Praat, we extracted the following vocal characteristics or these segments separated by speaker: (1) voice intensity, (2) fundamental frequency, (3) speech rate, and (4) first 12 MFCCs. Voice intensity is the loudness of the voice, which is measured in decibel (dB). The fundamental frequency is a physiological parameter defined as the frequency of vibration of the vocal folds, which can be measured in cycles per second or Hertz (Hz). The fundamental frequency is closely related to voice pitch, which is our human perception of fundamental frequency. The fundamental frequency generally lies in the range of 85–155 Hz for men and approximately one octave higher for women (165–255 Hz; for children, it is around 300 Hz). The speech rate or speaking rate is the tempo, in terms of the pace at which a stretch of connected discourse is delivered by a speaker. The average speech rate in conversations is typically around 3–4 words per second. We quantified the speech rate in vowels per second, using Praat software and the procedure described by De Jong and Wempe (2009). Finally, the first 12 MFCCs describe the spectral characteristics of an audio signal, in terms of different frequency bands that are determined logarithmically. They are commonly used to describe vocal behavior as their design is inspired by sound production in the human tract. These four vocal characteristics have been found to be effective for quantifying dyadic mimicry (e.g., Bonin et al., 2013; Solanki et al., 2016).

Extracting Motion Features From the Video Data

To capture motion from the video data, we extracted the thresholded pixel-wise difference between image frames in the region where a meeting participant is seated as a baseline feature. The more pixels change in intensity over consecutive frames, the higher the body motion score. Hung and Gatica-Perez (2010) used a similar approach to measure group-level cohesion except pixel-level movement was computed directly from compressed video data. The disadvantage of such approaches is that they are not able to capture movement or the shape of movement occurring over more than two consecutive video frames. Therefore, we also extracted dense trajectories (cf., Wang et al., 2013) which try to follow pixels (a 2×2-pixel neighborhood) over a pre-defined number of frames of a video, with a description per trajectory of the movement and appearance of the trajectory at each frame. Dense trajectories primarily describe the motion of a pixel in a video over time. The flow of the pixels is determined using optical flow fields computed on consecutive frames of the video. Then, the position of each location of the image in the next frame is determined based on the estimated flow direction. They are dense because the representation is extracted at multiple spatial scales, leading to a dense and overcomplete representation of the movement in the scene. This representation is able to describe human movement in an extremely powerful way for machine perception tasks. The default parameters are a trajectory length of L = 15 frames (see Wang et al., 2013), computed over an N × N pixels space-time volume aligned with a trajectory. The volume is subdivided into a grid, with

$$\text{grid size} = \mathrm{n}_\sigma * \mathrm{n}_\sigma * \mathrm{n}_\tau$$

The parameters n_σ and n_τ are used to specify the grid size of each space-time volume, where n_σ specifies the grid size in pixels in space and n_τ specifies the temporal grid size. Values for n_σ and n_τ can be 1 or higher, with previous work showing no superior performance beyond $\mathrm{n}_\sigma = 2$ and $\mathrm{n}_\tau = 3$ (Wang et al., 2013). We used the default parameters N = 32, $\mathrm{n}_\sigma = 2$, and $\mathrm{n}_\tau = 3$, which were previously found to perform well on action recognition tasks. Note that, in practice, the parameters tend not to be tuned but are determined based on empirical evidence. The identified trajectories can be highly non-linear in shape. The popular approach by Wang et al. (2013) which we also adopted in this paper does not regress the trajectory from the optical flow fields, so it is known to accumulate significant errors when longer trajectories are used. More powerful models can use more complex particle advection strategies to regress sub-pixel resolution trajectories, though this is more expensive to compute.

From the video data, we further extracted the HOG, a feature descriptor that is commonly used for object detection in computer vision and image processing (e.g., Zhu et al., 2006). The HOG represents the gradient directions of image intensity values as a distribution. These features provide a proxy of postural appearance over detected trajectories. Trajectories are associated with each person depending on whether the starting point of the trajectory originates within a pre-defined bounding box assigned for a given individual. We use a pre-defined bounding box in the image plane which represented a rectangular region over which the features for a given individual's behavior are extracted. In practice, we would prefer to be able to identify the seated person over time automatically. However, due to the uncertainty that the person would be tracked correctly and the knowledge in any case that the participants are seated throughout the meeting, we used a pragmatic approach to identify the relevant region of interest for each participant.

Extracting Movement Features From the Accelerometer Data

Accelerometer data captures movement along three orthogonal axes *X*, *Y*, and *Z*. The *XYZ* data represent the magnitude and direction of acceleration incident on each of the *X*-, *Y*-, and *Z*-axis, respectively. Because there can be interpersonal differences in the amount of movement in raw accelerometer signals, we first standardized each axis using *z*-scores. We then used these standardized values in three ways: the *z*-values themselves, the absolute values, and the magnitude which combines all three axes. We therefore obtained seven different features from the accelerometer data, namely, the raw and absolute value for each axis *X*, *Y*, and *Z*, respectively, and the magnitude for all axes combined.

We calculated absolute values, which removes the within-axis direction. The reason for using absolute values is that sometimes the direction of the movement along a particular axis might not be relevant (e.g., distinguishing forward from backward movement or left movement and right movement might not be necessary). The magnitude takes this a step further by considering just the amount of movement to be important irrespective of any direction. It is a representation of the total amount of movement a participant made and is calculated as

$$\text{magnitude} = \sqrt{X^2 + Y^2 + Z^2}$$

For accelerometer-based features, the output of the pre-processing step can generally be used as individual features. Kapcak et al. (2019) also propose using statistical and spectral representations of this data with a sliding window approach in order to capture variations in the body movements over time. This sliding window has a size of *n* and shifts by *n*/2 such that consecutive windows are half-overlapping. The size n of the sliding window is determined by trying varying window sizes and selecting the size that maximizes cohesion prediction performance on data not used for testing the approach.

Sliding Windows. Variables such as sliding window size or sample interval length are considered parameters in the machine learning model. Using a grid search approach, one can test various values of these parameters over a range of possible values to see which yields the best results on some data used for training the machine learner while then testing the resultant tuned model on new unseen data. Our sliding window size was a parameter that was empirically validated by using a simple toy model using the accelerometer data. Both the window size and sample interval length were fixed; in this case, the sample interval length was taken from prior work which had used the same thin slice length. However, parameters related to the weights for each of the coefficients of the logistic regressor were tuned for each of the five folds of the group-k-fold cross-validation. A logistic regressor has 1 + the number of input feature dimensions (in our case 720).

We tried sliding windows of 1, 3, 5, and 10 s (cf., Kapcak et al., 2019) and found that a 3-s sliding window yielded the best prediction performance. We compared each possible pair of participants in the team meeting using these 3-s sub-windows over the entire 2-min meeting slice (see Figure 4) and calculated the minimum, maximum, mean, and variance of the resulting differences as a representation of short-term similarity, as detailed next.

Computing Group-Level Mimicry and Convergence Features. To quantify group-level social signal mimicry and convergence for each of the included modalities, we computed group-level measures of mimicry and convergence based on measuring each person A's potential reaction to B in all possible dyads within each group. Hence, we first created pairwise similarity values for each dyad in the respective group and then aggregated these to the group level using the minimum, median, and maximum of standard deviation of the set of possible pairwise values (cf., Kapcak et al., 2019).

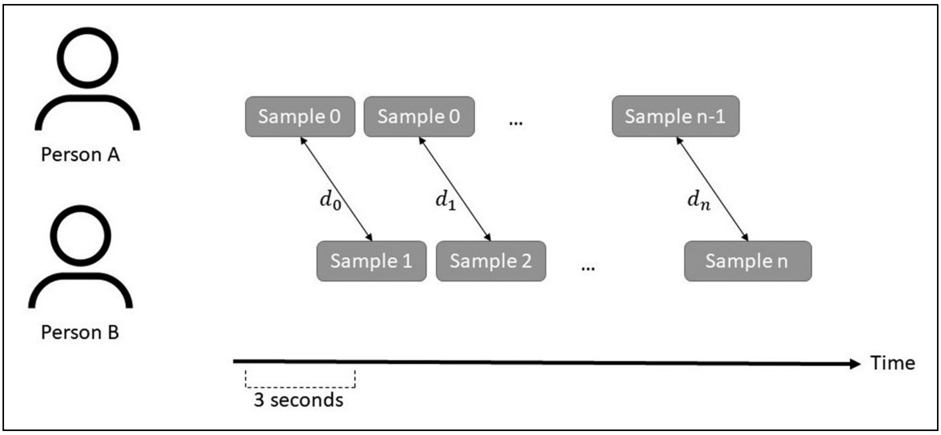

Figure 4. Computation of mimicry among team members. When person A is mimicking person B, they are imitating B's behavior with a delay. To measure short-term mimicry for each possible dyad in each meeting segment, we compared the two persons' consecutive 3-s sliding windows (shifting by n/2 such that consecutive windows are half overlapping, as illustrated) by calculating the distance between these windows (figure adapted from Kapcak et al., 2019). Across all of these possible comparisons within each 2-min meeting segment, we calculated the minimum, maximum, mean, and variance as a representation.

We then added this to the feature representation by placing the set of short-term similarity features corresponding with the highest maximum first.

Mimicry of the audio features was represented by the extent to which additional data samples from one speaker followed the distributions from the other speakers in the group. We implemented model-based similarity measures for the audio signals based on Nanninga et al. (2017) to compare paralinguistic mimicry features between participants, based on the log-likelihood as follows:

$$r_{ij} = \frac{1}{M} \sum_{i=1}^{N} \log(P(X_j|\theta_i)$$

where M is the number of data points of the speech signals extracted for participant j and θ_i are the parameters of the statistical model that characterize the speech of a different speaker i. We calculated the log-likelihood rather than the likelihood because the latter would result in very small numbers for some of the audio features. A better model fit, described by the average log-likelihood, indicates more similar speech behavior among each pair of meeting participants compared per corresponding sub-window.

From the set of dense trajectories (see overview in Figure 3) extracted for each participant delimited by a pre-defined bounding box in the image plane, we created a model for each participant in each meeting section. We adopted a Gaussian Mixture Model to learn a compact representation of each participant's dense trajectories in order to model their synchrony and convergence with other participants in the meeting (see Nanninga et al., 2017 for a similar approach for audio data, as well as Solanki et al., 2016, for more detail on using Gaussian Mixture Models to capture social signal mimicry). By applying the features obtained from another participant on this trained Gaussian Mixture Model, we could see how well they matched the movements of a given participant in the meeting.

For the accelerometer and video-based temporal difference (TD) data, relatively few dimensions were extracted over time. We therefore used five different forms of synchrony and convergence measures: normalized mutual information, Pearson correlation, short-term similarity (i.e., mimicry), global convergence, and symmetric convergence. Normalized mutual information is calculated as $I(X; Y) = H(X) + H(Y) - H(X, Y)$, where $H(X)$ and $H(Y)$ are the marginal entropy for both variables and $H(X, Y)$ is the joint entropy. In our case, this metric quantifies the amount of information that can be obtained about one participant's behavior by observing another participant in the meeting. From this, the normalized mutual information is calculated as follows, which yields a score between 0 and 1. A higher score means that the behavior of the two meeting participants is more similar:

$$NMI(X; Y) = \frac{I(X; Y)}{\sqrt{H(X)H(Y)}}$$

Pearson's correlations among the features from different attendees can range between -1 and 1, where a value closer to 1 means that the extracted features of two participants within the team meeting are more linearly related and are therefore more similar.

Short-term similarity compares the difference between feature values of one meeting participant A in a given time window to those feature values of participant B in the consecutive time window (see Figure 4). Note that our measure for short-term similarity or mimicry is asymmetric, because it yields different values when comparing participant A to B versus comparing participant B to A. Because it should not matter which participant is considered as person A and which one as person B, we added the mimicry comparison with the highest maximum to the feature vector first, followed by the one with the lower maximum.

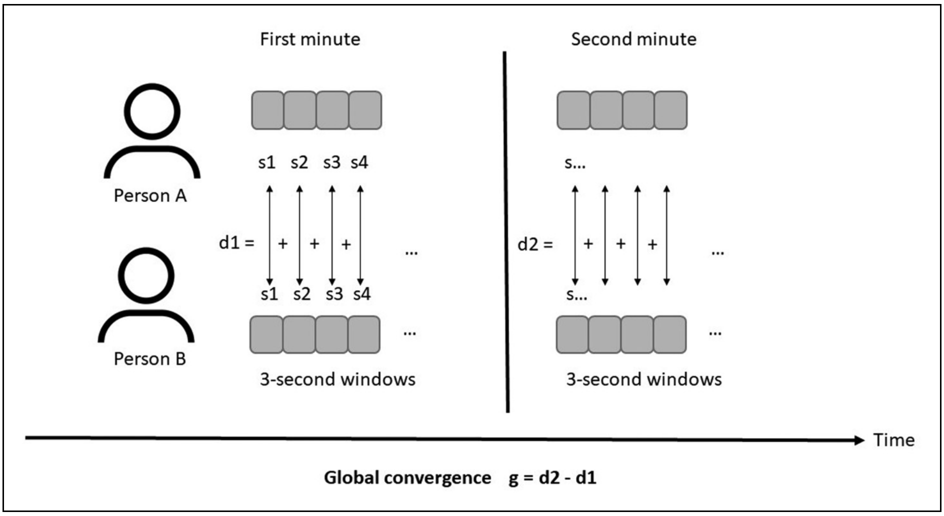

Figure 5. Calculation of the global convergence value. The distance between participants for each of the overlapping sliding windows (s) was summed for the first half of the 2-min segment as well as the second half of the segment, respectively.

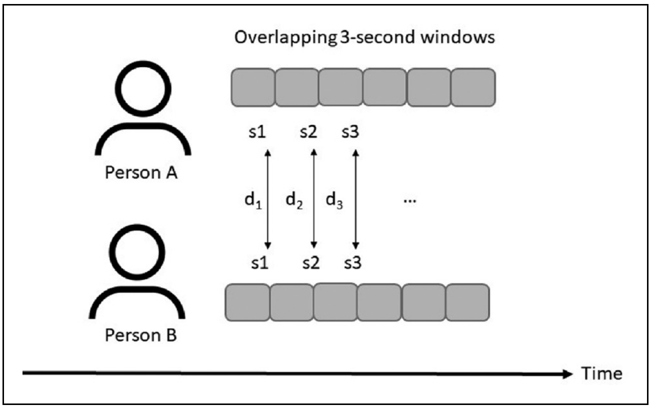

Figure 6. Calculation of symmetric convergence, in terms of the difference between feature values of two participants for each overlapping sub-window (3 s each). The correlation between these difference values over time is the symmetric convergence rate.

Global convergence is the difference in the behavioral features of two participants when comparing their behavior in the first half of each 2-min segment, compared to the second half of each segment (see Figure 5). Symmetric convergence is the correlation between features of person 1 and 2 in corresponding 3-s sub-windows across each 2-min segment (see Figure 6).

From each of these, four group-level measures were created by generating the minimum, maximum, median, and standard deviation from all possible pairs for each group for a given meeting segment. This led to $5 \times 4 \times (4 \times (6 + 2))$ features for the accelerometer data and $5 \times 4 \times 1$ features for the video-based TDs.

Synchrony was computed as the average likelihood of the samples from another participant belonging to a given person's model. Convergence was captured as pairwise similarity by computing the Pearson correlation per timestep and the likelihood. The same four group-level features were computed per pair as above, leading to 3 × 4 features for the audio and 4 × 3 × 4 features for the video-based HOGs extracted from the video data. We therefore extracted 720 features in total to train our algorithm.

Algorithm Development

Our aim was to create a machine learning algorithm that would be able to automatically detect the level of task and social cohesion (high or low) in each meeting segment, based on group-level social signal mimicry. If such an algorithm can indeed predict cohesion labels, then we can eventually use it on other datasets and avoid the hassle of laborious human annotation of team behavior. Instead, the algorithm will tell us about behavioral expressions of task and social cohesion in the observed team interactions. To get there, we needed to specify how the input from the various sensors should be processed and combined to automatically detect task and social cohesion expressions in each meeting.

In machine learning, a classifier is an algorithm that can automatically categorize data into one or more classes, such as an email classifier that scans emails and categorizes them as "spam" or "not spam." In our case, the classes were "high task cohesion" versus "low task cohesion" and "high social cohesion" versus "low social cohesion" for each 2-min segment in each team meeting. We had two goals in our algorithm development. First, we wanted to obtain reliable predictions of task and social cohesion. Second, we wanted to be able to interpret which specific behaviors help predict task and social cohesion inside the meeting. We used a so-called simple classifier, which allowed us to look at the weights of each feature and to identify which features contribute more to successful classifications.

Specifically, we used a logistic regression model and chose a linear classifier to classify high and low cohesion using the group-level behavioral mimicry features described above. Borrowed from the field of statistics, machine learning approaches use logistic regression as a so-called supervised learning classification algorithm in order to predict the probability of a target variable (in our case, labels for high versus low task and social cohesion, respectively). If researchers want to use this approach, the nature of the dependent variable needs to be dichotomous, which means there can only be two possible classes (in our case, high or low, represented as 1 or 0 for each meeting segment). Mathematically, our logistic regression model predicts the probability of high cohesion as a function of the group-level mimicry features for each type of modality. For the multimodal version of the algorithm including all available modalities, this level of probability was averaged across the different modalities.

K-Fold Cross-Validation

Machine learning algorithm development requires training data on which the algorithm learns how to automatically detect the criterion of interest (in our case, task and social cohesion in each meeting segment) and test data for evaluating algorithm performance. To evaluate the generalizability of a proposed approach, researchers in social signal processing usually perform cross-validation. This involves partitioning the data into k folds where each fold is used to test the model while the remaining k-1 folds are used to train the parameters of the model. Each time a different fold is selected as the test fold, and then, a different model is being trained based on different training data. In our case, we also forced each fold to be stratified by group such that no data from the same group having the same meeting would appear in both the train and test folds.

Due to our field sample, we only had a limited number of meeting segment data available, particularly for training and testing the algorithm across all modalities (see Table 2). To address this, we used k-fold cross-validation and set k to 5. We divided our data into k meeting exclusive folds of which one was used as a test set and the others were used as the training set, respectively, for the machine learning algorithm. This was repeated k times so that each fold was taken as the test set once. This approach reduces over-fitting and therefore gives a more accurate estimation of performance on previously unseen data (Cawley & Talbot, 2010).

Findings

To evaluate the performance of our algorithm, we examined how well the quantified group-level mimicry within each extracted feature predicted instances of high or low task and social cohesion, respectively. Machine learning approaches measure this prediction or classification performance using metrics. In this case, due to the imbalance of the high and low cohesion samples, we used the average area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The ROC curve represents the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds. Compared to other performance metrics, the AUC provides the most complete picture of how well a model is performing because the ROC on which it is based shows the performance at all possible combinations of precision and recall for the given test dataset. Appendix B provides more detail regarding reasons for selecting the AUC as a performance metric as well as how the AUC is computed, along with more information about the correlation of the different features with task and social cohesion labels across our dataset.

In addition to the AUC scores that follow, Appendix C presents different performance metrics (precision, recall, and f-measure), based on post hoc additional analyses using the extracted features. Performance metrics such as the f-measure provide an intuitive idea about model performance if the ultimate goal is to understand the efficacy of a model at the time of application. Whereas the AUC metric considers different classification thresholds, the f-measure only provides an estimate of a model's performance at a single point on the precision–recall curve. The selection of this point on the curve depends on what a user needs for a given application. If both classes (positive and negative) are considered equally important, one would choose a model threshold that maximizes both precision and recall (f1-score) jointly. However, some applications may favor higher precision and lower recall for detecting the positive over the negative class. For the interested reader, we provide results for the f-measure where precision and recall are considered equally important (see Appendix C). We caution the reader that the aim of this paper is not to provide a software tool that can be downloaded and used as is.

Comparing Feature Performance With Respect to AUC Values

The boxplot in Figure 7 shows the algorithm's classification performance for automatically detecting task cohesion (light gray) and social cohesion (dark gray) using the different features. This classification was based on the total of 720 features extracted across the entire dataset (as illustrated in Figure 2). As can be seen in Figure 7, all of the features performed better for predicting social cohesion than for task cohesion. Social cohesion was detected with an average AUC of 0.64 based on the accelerometer features, 0.63 based on the audio features, and 0.57 based on the TD features. Task cohesion was detected with a noticeably lower average AUC of 0.57 based on the accelerometer features, 0.60 based on the audio features, and 0.52 based on the TD features.

The reason why the audio features performed best for detecting task cohesion was likely that all team meetings had audio data, so the audio features had the largest amount of training data available (see Table 2). Because a different set of meetings was available for each modality, we cannot definitively conclude which modality detects social and task cohesion the best. Still, it appears that paralinguistic mimicry is a decent detector of both social and task cohesion, whereas motion similarity extracted from an accelerometer and TD data works for detecting social cohesion. The appearance-based HOG features seemed to detect neither.

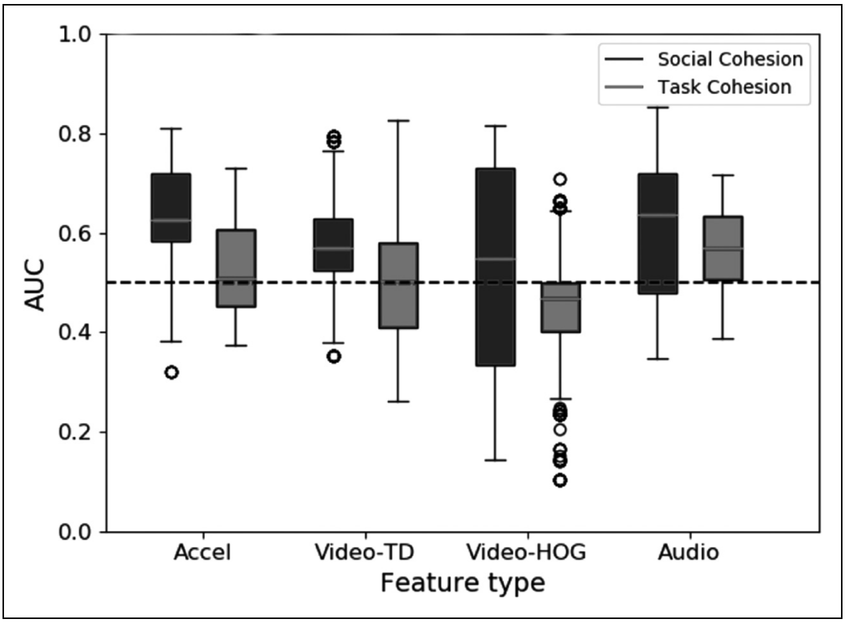

Figure 7. Boxplot of individual feature performance for predicting task cohesion (light gray) and social cohesion (dark gray) by each feature type, including error bars.
Note. AUC = area under the curve; TD = temporal difference; HOG = Histogram of Oriented Gradient.

To compare the predictive performance of the different features, we also tested the classification performance of each individual feature for the shared set (see Figure 8a). For the video features as well as the audio features, we obtained a better AUC for social cohesion than for task cohesion, respectively (see Figure 8a). Overall, the accelerometer features yielded the best performance for both social and task cohesion. Interestingly, the video TD approach captures a proxy of the speed of the movement rather than the acceleration, which may make the representation potentially more descriptive of the actual underlying behavioral process. The audio features performed poorly for detecting cohesion, compared to earlier analyses (Nanninga et al., 2017), because only a subset of our data had all modalities available for training and testing. Comparable performance was obtained when each modality was evaluated independently using all available data.

We also examined multimodal classification performance, which means that we let the algorithm combine different features to automatically predict task and social cohesion (Figure 8b). Note that we chose not to include the HOG features in the multimodal analysis, given that these features performed poorly for predicting either type of cohesion (see Figure 8a and Appendix B for more information). As depicted in Figure 8b, the combination of audio and TD features yielded the best average performance for the two modalities, and the combination of all three modalities (video, audio, and accelerometer) yielded the highest AUC. Whereas the unimodal performance of the audio and TD video features for detecting task cohesion was not that high, the combination of audio and video led to an average AUC of 0.62. In comparison, combining the accelerometer and TD features did not improve over the performance of the accelerometer data alone—which is perhaps less surprising given that both measure movement similarity.

We obtained the best average classification performance when combining motion-based mimicry features from the accelerometer and video data with the audio features, which suggests that the

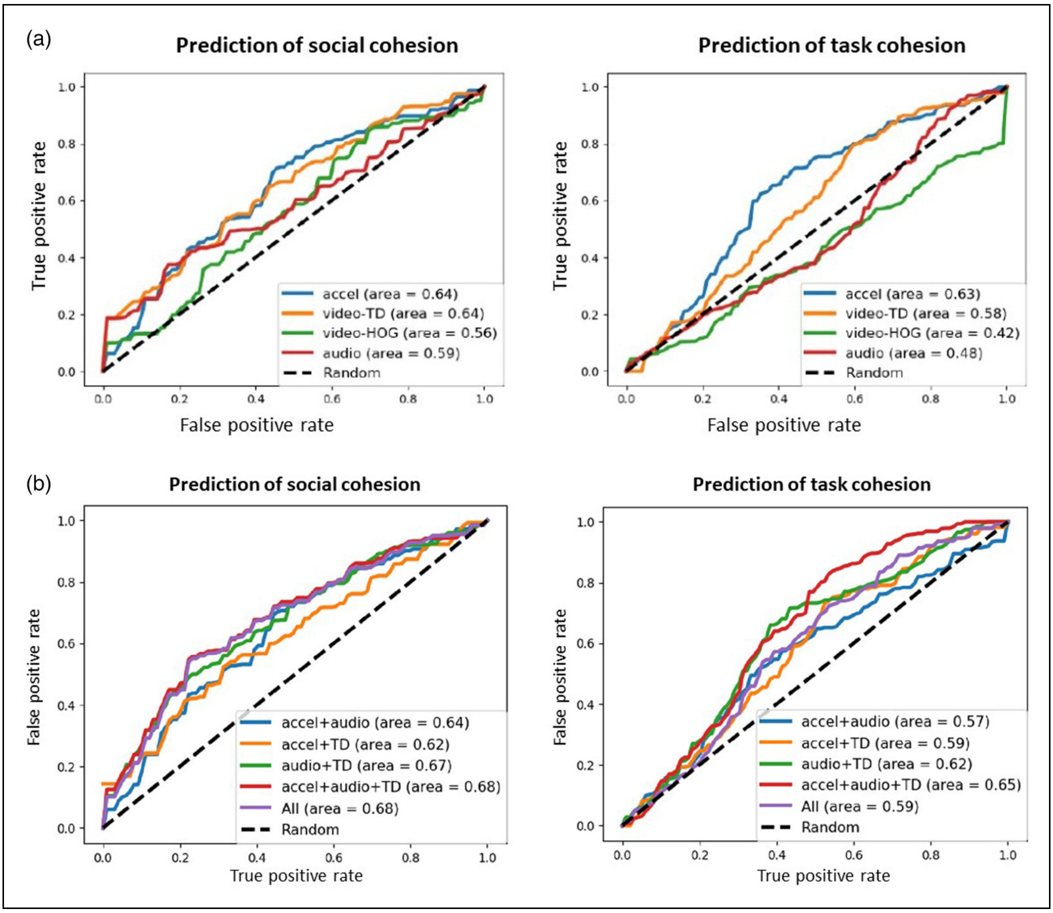

Figure 8. (a) Quality of the automated prediction of social cohesion (left) and task cohesion (right) by each individual feature type, as indicated by the respective average ROC curve. (b) Average ROC curve for multimodal combinations of the features in order to automatically detect social cohesion (left) and task cohesion (right), respectively.

Note: ROC = receiver operating characteristic; TD = temporal difference; HOG = Histogram of Oriented Gradient.

combination of all different modalities does contain complementary information. However, the increase in performance was just 0.01 for social cohesion when compared to the combination of audio and TDs. Hence, while we cannot conclude that the combination of all modalities provides a significant increase in average AUC for social cohesion, the combination of features representing the two different mimicry types does seem beneficial for the detection of social cohesion. Similarly, when combining all modalities to detect task cohesion, the performance was higher compared to the detection based on accelerometer features alone, which suggests that a multimodal model detects task cohesion better than a unimodal model.

In sum, this study showcases the potential of social signal processing for automatically detecting behavioral manifestations of team constructs such as cohesion. Our findings illustrate how a multimodal approach, combining audio, movement, and video sensor data that captures mimicry at the group level, achieves the most accuracy for automatically predicting instances of social cohesion. Task cohesion was generally more difficult to predict automatically, which may partially be due to the higher overall frequency of meeting segments labeled with social cohesion compared to task cohesion behaviors in our dataset.

Discussion

Interdisciplinary work can bridge social and computer science in general and innovate methodological approaches to team interaction dynamics specifically. First, our study illustrates how a social signal processing approach can provide novel, unobtrusive measures of highly granular team interaction behaviors. To this end, we want to emphasize that automatically detecting behavioral team constructs such as cohesion still requires substantial input from social scientists in order to develop behavioral measures and establish the "ground truth" that serves as the basis for social signal processing and machine learning. In other words, as tempting as it may seem, the automated detection of team phenomena without any human annotation effort is not yet possible. To get there, more interdisciplinary collaborations between social and computer scientists are needed to create data available for computer scientists to train more robust models in "messy," real-life team interaction settings such as the on-site project meetings investigated here.

Second, our interdisciplinary empirical approach offers a potential solution for overcoming the limitations inherent in static, survey-based measurement approaches to team constructs (e.g., Kolbe & Boos, 2019; Kozlowski & Chao, 2012; Lehmann-Willenbrock & Allen, 2018). Considering the benefits of surveys regarding the ease of data access and the low effort involved in processing and analyzing the data, compared to behavioral observations of team processes, the departure from survey-based designs can seem daunting. We hope that our study will encourage team researchers to seek out interdisciplinary collaborations that can facilitate this leap and advance our understanding of the intricate behavioral dynamics at the core of complex team phenomena such as cohesion. We suggest that team researchers should consider incorporating multiple sensors in their study design, including movement as well as audio and video data collection. The methodological setup described in our study is one example of implementing such an approach without interfering with the naturally occurring team interaction, while still obtaining multimodal behavioral data from the field.

Third, in showing how behavioral indicators of task and social cohesion in real-life team meetings are represented by the dynamics of social signals at the micro-level and across different behavioral modalities, our findings point to the importance of capturing behavioral mimicry in team interactions. Our finding that a combination of team-level movement and paralinguistic features performed best for automatically predicting behavioral team cohesion within the meeting suggests that mimicry in teams is a multimodal phenomenon. This insight speaks to calls for multimodal approaches in social signal processing more broadly (Baltrušaitis et al., 2018). An important benefit of the social signal processing approach is the ease with which multimodal behavioral phenomena can be investigated. In our case, we only looked at within-modality behavioral synchrony. Additional analyses could be carried out involving cross-modal synchrony measures, which would allow us to probe questions related to the role of inter- versus intra-modal synchrony for predicting team constructs. Furthermore, in addition to synchrony, other higher-order terms worth exploring include possible interactions between modalities (e.g., specific audio and video features might interact to predict different types of cohesion differently).

Fourth, beyond our specific example of predicting team cohesion, our findings suggest that social signal processing can be used to automatically detect complex social interaction phenomena more broadly. Indeed, computer scientists have applied social signal processing to a broad range of constructs of relevance to organizational researchers, such as detecting emergent leadership influence (e.g., Muller & Bulling, 2019), predicting group performance from social signals (e.g., Murray & Oertel, 2018), or automated analyses of group affect (e.g., Böck, 2021). Gatica-Perez and colleagues (2017) reviewed 100 social signal processing publications that focused on small group analysis and derived conversational dynamics, verticality (including dominance and leadership), team personality, and group characterization as trending topics in social signal processing that are of high relevance for organizational researchers interested in group and team dynamics, leader–follower dynamics, and social interactions at work more generally. Moving beyond the typical controlled laboratory setting in these earlier applications, our findings from real, "messy" team meetings in organizational context show that multimodal social signal processing approaches can be extended successfully to teams in the wild. As a caveat, we need to acknowledge that we are not in a position yet to fully rely on machine learning algorithms to reliably detect complex team behavior.

Our proposed approach can be easily extended to include the processing of other modalities including language, for example, by processing the transcripts generated from automatic speech recognition using pre-trained neural language embeddings. We could indeed also consider processing other information at the utterance level. However, there are privacy concerns to consider. Because we gathered field data from real project meetings, verbal transcriptions of the meetings were strictly forbidden. Given this practical constraint, it is worth considering the benefit of also developing multimodal nonverbal approaches further. This privacy-preserving aspect of our processing approach was certainly appealing to the company where we gathered the data and was a contributing factor to our collaboration with them.

Limitations and Future Research Directions

Several limitations of our approach point to future research opportunities. Some of these concern our choice of labels for dynamic cohesion. In our machine learning model, the level of cohesion was binarized. The choice to simplify the data by creating labels for high versus low task and social cohesion, rather than using continuous labels, is standard in the social signal processing community, yet this leads to a loss of detail. It remains a research challenge for the social signal processing community how to best handle these data which have neither positive nor negative classifications. One might be led to think that solving this with a regression task would be the solution. However, given the skewed nature of the data distributions as shown in Appendix A, Figure A1, computing regressions accurately remains an open problem in the field. Moreover, metrics for measuring good performance on a regressive scale tend to be harder to link back to what is actually happening in the social interaction (in our case, a meeting). This remains an open design choice for any interdisciplinary collaboration in the area of social signal processing.

Moreover, our machine learning model considered the two cohesion estimation tasks independently. In other words, all 2-min meeting segments were labeled as either task or socially cohesive, but never as both. Given that each time window within the team interaction stream could contain both utterances of task and social cohesion, each window could receive a continuous value label for the associated level of task and social cohesion present, rather than the binary categorization into high/low cohesion as applied in our case study. For instance, each behavioral unit within the team interaction would be annotated with a certain percentage of task cohesion and social cohesion. In adjusting the outputs of the estimation task and considering the problem to be one of jointly estimating task and social cohesion, other machine learning approaches such as multitask learning could be used to better predict both types of cohesion with the assumption that there is some dependency between the two phenomena that can also be learned. Notably, much more training data would be required to achieve reliable predictions on continuous labels of task and cohesion levels inside meeting segments.

Our choice of logistic regression as a supervised learning classification algorithm was one possibility among many, many others. It really remains an open question which approach would be most accurate. One of the issues to be considered in the context of our current study is that different partitions of the data would yield different results, given the type of validation approach we used due to the small sample. It would be beyond the aims and scope of this paper to compare different classifiers, but we refer the interested reader to Witten et al. (2005) for an overview of different types of algorithms and data mining methods in machine learning. Moreover, given the growing success of deep learning approaches, future research might explore machine learning approaches that learn features for social interaction from data (i.e., learned representations) rather than using prior knowledge to design features. Such deep learning approaches use artificial neural networks (e.g., Rosopa, 2022; Urban & Gates, 2021) and rely on massive amounts of data, which are far beyond the amounts available from typical samples of behavioral team interactions such as the one examined in the current study. Likewise, leveraging knowledge that can be generalized from multiple datasets, recording conditions, and contexts remains an open question. There is still a hope that with learned representations, learned models can still be adapted to a specific set of experimental data using adaptation approaches such as fine-tuning. However, future research using deep learning to detect behavioral team phenomena should aim for larger samples of in-the-wild team interactions than the one we gathered and annotated here. Moreover, we caution the reader that while deep learning approaches are powerful machine learning techniques, they fall short in terms of their interpretability. It remains an open question as to how we can resolve the semantic interpretation of such abstract representations, which can often be "in the eye of the beholder." This phenomenon is often overlooked and can cause severe misinterpretations of machine learning results if not handled appropriately (see also Raman et al., 2022, for a discussion of this point).

In the social signal processing field, works that are showing proofs of concept tend to use simpler classifiers with feature engineering, while contributions that are more technical typically propose more robust classifiers that are designed based on hypotheses about the nature of the prediction problem. While some earlier SSP works did compare modeling with different classifiers, we deem this a rather unsatisfactory approach. This is because the explanation for why one classifier works better than the other cannot be meaningfully extracted from more complex classifiers without further analysis and often does not lead to satisfactory conclusions. This is further compounded by the comparatively small size of our field study dataset (compared to typical dataset sizes in the machine learning field). By using more sophisticated machine learning classifiers, we would reduce the model transparency and might overfit our model to unwanted nuances in the data. How to trade off these issues remains an open question for interdisciplinary research.

In the scope of modeling choices, future research could also examine to what extent data transformations may improve performance in social signal processing. Methods from generalized linear models (e.g., Rönkkö et al., 2022), such as log transformation, could potentially be applied to machine learning model performance when labels are not distributed evenly (as in the case of our labels for high and low cohesion, see Table A1 in Appendix A). Moreover, future research can investigate the potential interplay between micro-level behavioral mechanisms underlying task and social cohesion, as presently investigated, and larger-scale temporal dynamics of (emergent) task and social cohesion across longer interaction periods, such as entire meetings or series of meetings by utilizing multilevel modeling approaches.

Regarding the ground truth for variables such as cohesion at the behavioral event level, we took some liberties in our case study to establish a link between verbal utterances and labels for task and social cohesion. We provided conceptual arguments for this link, drawing from the extant literature on static measures of cohesion. Of note, we do not claim that our labels of behavioral-level cohesion are the same as overall emergent cohesion, which is typically measured using self-report surveys (e.g., Salas et al., 2015). Equating these different types of variables would be problematic, for example, because survey-type reports of a very spontaneous, subconscious coordination process such as the behavioral mimicry of social signals that represent moments of cohesion during team interactions would not be feasible. In other words, we cannot expect team members to self-report fluctuations in cohesion at the micro-level (see Mayo & Gordon, 2020). It would be like trying to ask people to annotate the mimicry while experiencing the interaction. We might label mimicry from external perceptions, but actually, we already know that these kinds of nonconscious behavioral patterns are best captured using sensors (e.g., Vinciarelli et al., 2011).

Related to our operationalization of ground truth labels for behavioral cohesion, future research can investigate questions related to construct validity when pursuing a social signal processing approach to team interactions in meetings and other dynamic social settings. Construct validity remains a potential concern regarding the work we presented here, as it pertains to the operationalization of behavioral-level team cohesion, our key phenomenon of interest and the criterion which we used to train and evaluate our machine learning algorithm. We anticipate that it will be a research challenge in and of itself to examine whether and how micro-level fluctuations in behavioral expressions of cohesion map onto team members' self-reported experiences of team cohesion (which are prone to known biases associated with self-report methods but nevertheless insightful because they add a perceptual measurement layer; e.g., Gerpott et al., 2020). While beyond the scope of the current paper, future research can and should extend our approach to other team constructs that have not yet been operationalized at the behavioral level such that machine learning approaches can become more mainstream and more researchers can benefit from them. Moreover, future research can examine to what extent automatically detected behavioral expressions of task and social cohesion predict objective team performance outcomes, in light of previously established linkages between cohesion and performance in the survey-based literature (for an overview, see Grossman et al., 2022).

In our research example, regardless of the specific behavioral modality, the AUC score as an indicator of the quality of the automated prediction was not particularly high. Hence, the goal to replace laborious human annotations completely with artificial intelligence and rely on machine learning algorithms to tell us how teams "tick" is still a work in progress. Ideally, future work in this area will examine the performance of social signal processing approaches to group cohesion across a range of different field settings beyond the one examined in our study, thus further contributing answers to calls for more research "in the wild" voiced in both disciplines (e.g., Alameda-Pineda et al., 2018; Shuffler & Cronin, 2019). To advance interdisciplinary research in this direction, more collaborations between social scientists and computer scientists are needed. This includes the need for collaborative datasets, along with new policies and ethical procedures in place for sharing such data, and interdisciplinary foresight when designing studies (Keyton & Heylen, 2017). For researchers interested in embarking on such projects, Table 3 lists key points for consideration during research design in order to enable multimodal social signal processing.

Finally, the key term "ground truth" in social signal processing warrants some reflection. For organizational researchers focusing on team processes, the idea of a "truth" can seem rather odd, given the subjectivity of team experiences and the fact that team research is not an exact science. As a side note, while machine learning researchers call the labels "ground truth," others such as the speech processing community call them the "reference." This accounts for the fact that the labels themselves might be subjective in some way. Historically, many traditional machine learning tasks were able to offer a more objective (true) idea of what a label should be. However, as the field has developed and more subjective phenomena are being investigated, the use of soft labels or the notion of subjectivity in the labeling process is also accepted. However, the way of modeling this subjectivity in machine learning approaches remains an open question.

To this end, scholars need to invest in developing and validating behavioral coding schemes for behavioral team constructs such as task and social cohesion. As a first step, team phenomena need to operationalize team constructs at the behavioral level such that they become observable. More scholars need to focus on team dynamics in an organizational context with methods that yield high-resolution team interaction data (e.g., Klonek et al., 2019). Organizational researchers should seek dialogue with computer scientists in order to establish data-gathering routines, including the requirements in the observational setup to gather relevant social signals at sufficient quality and enable

Table 3. Research Design Aspects to Enable Multimodal Social Signal Processing.

Design Aspect	Points for Consideration
Location	• Ideally, no windows; no changes in ambient lighting • Ideally, soundproof or in a quiet building area to avoid outside noise on the recorded audio • No fans/ventilators/noisy air conditioning noises inside the room • Well-illuminated room to avoid shadows on participants' faces
Sensor selection and setup	• Individual videocameras trained on each participant (when seated, face and upper body should be centered in each video). Camera signals with high resolution at a constant frame rate (in our study, 1,920 × 1,080 recorded at 25 fps) • Individual microphones (headsets preferable over lapel microphones). Headsets should not be worn too close to the mouth to avoid signal saturation, e.g., from the utterance of plosives or sudden laughter • Range of motion tracking technology (e.g., Poppe, 2011), including obtrusive (e.g., accelerometer badges) and unobtrusive (e.g., depth cameras)
Feature alignment	• Necessary to temporally align data obtained from different participants and from different sensors (audio, video, motion tracking) • For video data: participants' actions are compared in time, so the frames that occurred at approximately the same timestamp within different videos need to be found. This requires sampling video data at a constant frame rate. In our case, frames with their timestamps closest to the new timestamp were used to construct the new video using ffmpeg • Video/audio synchronization (e.g., Lichtenauer et al., 2011; Raman et al., 2020): ideally, digital synchronization. Manual synchronization, comparing statements heard in the audio with visible mouth movement in the video, is less preferable because it can cause multimodal drift • Synchronizing accelerometer with audio/video data requires a precise data gathering protocol, e.g., in our case, a research assistant was asked to clap a specific accelerometer on the table after reading out the time shown on that accelerometer

robust multimodal social signal processing, and appropriate time windows for annotating behavioral team constructs. Alternatively, interdisciplinary research teams can explore to what extent the raw signals from sensor data can provide meaningful insights into team constructs directly. The larger question in this regard concerns the notion of representation, that is, how we may build scholarly consensus to move beyond abstract sensor information and toward meaningful social signals that represent psychological phenomena in and of themselves.

Implications for Interdisciplinary Collaboration in Social Signal Processing

Our mode of collaboration in this study can be considered a light form of the producer–consumer model (see Allen et al., 2017), where the social scientist provided the data and human annotations of the observed team interactions and advised the selection of the task and socially cohesive utterances and final interpretations of the resultant developed model. While such a lighter collaborative construction can help in the earlier phases of establishing a common ground for interdisciplinary communication, one would hope for more developed findings and outcomes moving forward with this collaboration. Looking back on the collaboration, we could have taken a more equal role with respect to the generation of joint research questions and the social scientist could have been included more in the design decisions regarding the creation of the extracted feature data. This could have perhaps led to more application-driven motivations for different machine learning models such as multitask learning or a different labeling approach based on the proportion of utterances that were task or socially cohesive within each 2-min meeting segment, as discussed above. Despite the early stage nature of this work, the joint experience has been invaluable as a common reference point to develop a joint working language for the collaboration and for furthering interdisciplinary research collaboration.

Regarding the practical implications of how such interdisciplinary work can be carried out, one of the biggest hurdles is typically related to the lack of funding, which can be a chicken-and-egg problem (i.e., without funding, no work can be carried out, and without at least preliminary work, there is typically no funding). Our experience shows that this is a common misconception and that actually low-hanging fruits do exist. While we conceptualized the research idea and study design and set up the collaboration with the organization where the data were gathered, the actual data gathering and much of the analytical work presented in this paper were then carried out by graduate students in an interdisciplinary team as part of their master theses. One format could involve one or more graduate students from each discipline working together. Beyond coordinating such an effort, this also requires that the PIs are willing to invest time in advising graduate students from the respective other discipline. As a caveat, such an approach limits the level of transdisciplinarity that the work can have since graduate students need to fulfill certain disciplinary requirements. However, such a collaboration forms a basis for establishing a common ground and language. From there, we can develop more mature co-designed research questions that can ultimately advance our conceptual understanding of social dynamics in groups and teams and computing systems that can support organizations in improving teamwork. In closing, we want to emphasize the exciting research opportunities afforded by social signal processing for innovating team science. At the same time, we are nowhere near replacing human annotators with machines in order to understand complex team phenomena such as cohesion. Continued collaboration and potentially data sharing with the wider research community will allow computer scientists to develop more robust automated machine perception approaches toward understanding more complex behavioral group and team phenomena. We hope that the growing interest in interdisciplinary initiatives in this realm (see Hung et al., 2020, for a summary) will continue to build.

Appendix A: Distribution of ground truth labels across the data set

Table A1. Percentage of 2-minute Segments (total N = 600) Within Each Meeting Labeled as High or Low Social and Task Cohesion, Respectively.

			Percentage of 2-minute Segments Labeled as:			
Meeting Number	Duration (hh:mm:ss)	Number of Labeled 2-Minute Segments	High Social Cohesion	Low Social Cohesion	High Task Cohesion	Low Task Cohesion
1	01:14:00	37	.541	.000	.081	.270
2	01:27:45	42	.238	.119	.357	.500
4	01:04:05	31	.290	.419	.226	.452
5	00:52:45	26	.385	.115	.538	.269
6	01:29:41	44	.091	.636	.023	.841
7	00:41:23	20	.250	.200	.000	.750
8	00:22:29	11	.091	.091	.455	.091
9	00:24:30	12	.000	.667	.000	.583
10	00:55:40	27	.185	.481	.296	.370
11	00:47:13	23	.217	.348	.348	.217

(continued)

Table A1. (continued)

			Percentage of 2-minute Segments Labeled as:			
Meeting Number	Duration (hh:mm:ss)	Number of Labeled 2-Minute Segments	High Social Cohesion	Low Social Cohesion	High Task Cohesion	Low Task Cohesion
12	00:43:00	21	.429	.048	.238	.333
13	00:58:38	27	.259	.259	.000	.444
14	00:36:45	18	.167	.389	.333	.333
15	00:54:25	27	.259	.074	.074	.593
16	00:28:38	14	.786	.000	.000	1.00
18	00:41:29	18	.167	.333	.056	.444
19	01:13:20	36	.250	.139	.222	.139
20	00:42:55	21	.476	.000	.476	.048
21	00:49:16	24	.083	.417	.208	.458
22	00:49:13	24	.292	.125	.833	.042
23	01:01:10	30	.067	.133	.733	.000
24	01:11:24	36	.250	.139	.222	.139
25	01:02:28	31	.065	.516	.065	.613

Table A2. Means, Standard Deviations, and Correlations of Cohesion Labels (in %) Across the Entire Data Set.

	M	*SD*	High Social Cohesion	Low Social Cohesion	High Task Cohesion	Low Task Cohesion
High social cohesion	.254	.179	1	−.724*	.094	−.135
Low social cohesion	.246	.204		1	−.311	.104
High task cohesion	.252	.235			1	−.779*
Low task cohesion	.388	.264				1

Note: $N = 600$ windows labeled for high/low task and social cohesion across the 25 observed team meetings. Spearman Rho correlations. $*p < .01$.

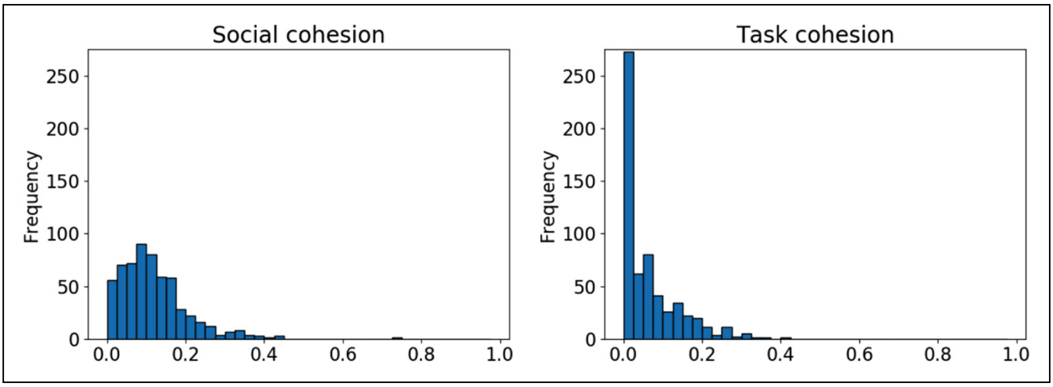

Figure A1. Frequency of behavioral expressions of social and task cohesion across the entire data set (i.e., all 2-minute meeting segments combined).

Appendix B: AUC calculation and individual feature performance

In social signal processing, it is not customary to report a range of different performance metrics because each metric should be well motivated with respect to the sample distribution. Given the highly imbalanced class sizes in our case (see Figure A1 in Appendix A), we chose the area under the curve (AUC) as a performance metric. The AUC metric is based on the receiver operating characteristic (ROC). In machine learning, the ROC is a popular diagnostic tool for evaluation the performance of classifiers both on balanced and imbalanced binary prediction problems because "ROC analysis does not have any bias toward models that perform well on the minority class at the expense of the majority class—a property that is quite attractive when dealing with imbalanced data" (Weiss, 2013, p. 27). Compared to other performance metrics, the AUC provides the most complete picture of how well a model is performing because the ROC on which it is based shows the performance at all possible combinations of precision and recall for the given test dataset.

Computing the AUC

The AUC is computed by first determining the ROC, as shown in Figure 8a and b in the paper. This shows how a model performs with respect to the true positive rate (TPR) against the false positive rate (FPR). This is achieved by computing the TPR and FPR at various different thresholds with respect to the probability of an observed sample belonging to the positive class (in this case, high cohesion). The computed TPR and FPR at each threshold can be plotted as a curve from which the area under it can be computed to obtain the AUC. One can imagine that the better our logistic regressor is at modeling the two classes, the more likely that the estimated probabilities reflect the true class label.

For the case of a perfect model, we can imagine that at the optimal threshold, all samples with a probability below that threshold should be correctly labeled as negative (low cohesion) and correspondingly, anything above that threshold should be correctly labeled as positive (high cohesion). At that moment, the ROC should trace a line going straight up the y axis until it reaches 1 and then straight across until it reaches 1 FPR and 1 TPR. We assumed in our case that positive labels refer to the high cohesion label and negative to low cohesion.

To compute the FPR, we first need to calculate the false positives (FP); the number of samples that were incorrectly labeled as having high cohesion, the true negatives (TN); the number of samples that were correctly labeled as having low cohesion. Then, we can compute FPR = FP/(FP+TN). In other words, of all the low cohesion samples, what percentage was labeled incorrectly?

To compute the TPR, we first need to calculate the true positives (TP); the number of samples that were correctly labeled as having high cohesion, and the false negatives (FN); the number of samples that were incorrectly labeled as low cohesion. Then we can compute the TPR=TP/(TP+FN). In other words, of all the high cohesion samples, what percentage were labeled correctly?

Note that to combine or fuse multiple modalities together in this study, we used a late fusion approach that takes an average of the output probabilities of each modality specific classifier. If we wanted to tune the performance, we could have combined the probabilities with an additional learned hyper parameter in order to better learn how to weigh the combination of different modalities together for both social and task cohesion.

Features most related to social and task cohesion (correlations)

The accelerometer features and the video temporal difference features showed the best performance (see Figure 8a). For more in-depth insights, we calculated correlations between the various features and task cohesion as well as social cohesion. Table B1 below shows the accelerometer and video-TD features which were most strongly correlated with the ground truth labels for social cohesion. Table B2 further below shows the accelerometer and video temporal difference features which were most strongly correlated with the ground truth labels for task cohesion. The first component in each feature name is the group-level aggregation method (e.g., Minimum), the second is the mimicry measurement (e.g., Mutual Information), and the last is the individual feature type (e.g., PSD_Z_6).

Both Table B1 and Table B2 show that the group minimum value is the most common among those features that correlate strongly with the cohesion labels in our data set. The group minimum value is always negatively correlated with the label (both for social and for task cohesion predictions). This may seem counterintuitive, because you might think that a group would be more cohesive when they are more similar. However, note that the focal speaker is not separated from the rest of the group. Hence, a low minimum results for example when the group is silently listening while the speaker shows more movement (which makes the movement pattern of the rest of the group very dissimilar to the speaker). This would also be reflected in a higher standard deviation, which is indeed positively correlated with both social and task cohesion.

Concerning social cohesion (Table B1), we observe positive correlations with the median and maximum group values among the video-TD features. This suggests that if the non-speakers of the group are more similar their social cohesion is also similar, because this would be reflected in a higher median and maximum. Mutual information is the pairwise correlation value that is most

Table B1. Correlations Between Accelerometer or Video-TD Features and Social Cohesion.

Feature Name	Pearson's r With Social Cohesion Labels	p-value
Accelerometer features		
Min Mutual Information PSD_Z_6	−.293	<.0001
Min Mutual Information PSD_X_3	−.292	<.0001
SD Normalized Mutual Information Var_XAbs	.291	<.0001
Min Normalized Mutual Information PSD_X_3	−.280	<.0001
SD Normalized Mutual Information Mean_XAbs	.267	.0001
SD Normalized Mutual Information PSD_Z_4	.262	.0001
SD Normalized Mutual Information PSD_X_2	.260	.0001
Min Normalized Mutual Information PSD_Z_6	−.251	.0002
SD Normalized Mutual Information PSD_X_1	.247	.0002
Min Mutual Information Var_X	−.246	.0002
Video temporal difference features		
Median Mutual Information Var_Mag	.220	.001
Median Mutual Information Mean_Mag	.186	.004
Max Mutual Information Var_Mag	.183	.004
Min Global Convergence Mean_Mag	−.181	.005
Min Mimicry_1_Mean Mean_Mag	−.164	.011
SD Global Convergence Mean_Mag	.162	.011
Min Mimicry_2_Mean Mean_Mag	−.161	.012
Max Mutual Information Mean_Mag	.161	.012
Min Mimicry_2_SD Mean_Mag	−.159	.013
Min Global Convergence Var_Mag	−.158	.014

Note: N = 600 meeting segments lasting 2 minutes each, labeled for high or low task and social cohesion. Min = minimum; Mag = magnitude; Max = maximum; PSD = power spectral density and the bin number; SD = standard deviation; Var = variance; X = *X*-axis; Y = *Y*-axis; Z = *Z*-axis.

Table B2. Correlations Between Accelerometer or Video-TD Features and Task Cohesion.

Feature Name	Pearson's *r* With Task Cohesion Labels	*p*-value
Accelerometer features		
Min Mimicry_1_Mean PSD_ZAbs_4	−.331	<.0001
Min Mimicry_2_Mean PSD_ZAbs_4	−.328	<.0001
Min Mimicry_2_Mean PSD_YAbs_5	−.314	<.0001
Min Mimicry_1_SD PSD_YAbs_5	−.313	<.0001
Min Mimicry_2_SD PSD_YAbs_5	−.311	<.0001
Min Mimicry_1_SD PSD_ZAbs_4	−.310	<.0001
Min Mimicry_1_M PSD_YAbs_5	−.308	<.0001
Min Mimicry_2_SD PS_ZAbs_4	−.306	<.0001
Min Mimicry_1_Max PSD_YAbs_5	−.298	<.0001
Min Mimicry_2_Max PSD_YAbs_5	. −.296	<.0001
Video temporal difference features		
SD Global convergence Var_Mag	.151	.009
Min Global convergence Var_Mag	−.134	.020
Min Mimicry_1_Mean Mean_Mag	−.133	.021
Min Mimicry_1_SD Mean_Mag	−.127	.027
Min Mimicry_2_SD Mean_Mag	−.126	.029
Min Mimicry_2_Mean Mean_Mag	−.125	.029
Min Correlation Var_Mag	−.125	.030
Min Correlation Mean_Mag	−.121	.036
Max Symmetric convergence Var_Mag	.120	.037
Min Mimicry_2_Max Mean_Mag	−.119	.039

Note: N = 600 meeting segments lasting 2 minutes each, labeled for high or low task and social cohesion. Min = minimum; Mag = magnitude; Max = maximum; PSD = power spectral density and the bin number; SD = standard deviation; Var = variance; X = *X*-axis; Y = *Y*-axis; Z = *Z*-axis.

commonly found in the top 10 correlation accelerometer features. Mutual information is a metric that quantifies how much about one signal can be known by knowing the other, and is therefore a similarity rather than a mimicry metric.

Table B2 shows that mimicry is a common entry both among the accelerometer features and among the video-TD features that correlate strongly with task cohesion.

Appendix C: Additional calculation of different performance metrics

In addition to the preferable AUC metric (see Appendix B for a detailed justification), alternative performance metrics were requested during the revision process. As this was not anticipated, we had not saved intermediate results that would enable alternative metrics to be directly calculated from the posterior probability per sample per repeated folder. Hence, we performed a separate analysis to obtain the requested alternative performance metrics based on the extracted features and used a so-called brute force approach to identify the combination of extracted feature files that would produce results with the closest AUC to our originally generated results. As an additional verification step, we cross-checked the ROCs and AUCs with our original findings and observed a mean difference of .078 across all AUC scores, which can be attributed to expected variations in folds of the cross-validations given our random seed approach to determine the data folds during our original analysis.

Table C1. Alternative performance metrics obtained from additional post-hoc analyses.

	Precision		Recall		F-measure (F1)	
	Social Cohesion	Task Cohesion	Social Cohesion	Task Cohesion	Social Cohesion	Task Cohesion
Accelerometer	55.5% (± 19.5%)	63.5% (± 14.0%)	51.9% (± 15.3%)	59.2% (± 12.6%)	50.4% (±17.6%)	58.7% (± 12.2%)
Video TD	65.0% (± 9.6%)	54.7% (± 12.4%)	55.4% (± 11.5%)	73.3% (± 8.3%)	52.8% (± 9.57%)	62.5% (± 11.2%)
Audio	44.7% (± 25.6%)	54.7% (± 12.4%)	54.3% (± 9.7%)	73.2% (± 8.2%)	43.7% (± 15.6%)	62.4% (± 11.2%)
Accelerometer + audio	55.7% (± 17.9%)	60.0% (± 13.0%)	50.7% (± 12.4%)	61.0% (± 13.0%)	48.6% (± 14.7%)	58.1% (± 13.6%)
Video TD + audio	60.0% (±18.4%)	54.8% (± 15.8%)	50.5% (± 7.9%)	73.5% (± 8.6%)	41.3% (± 6.5%)	62.6% (± 11.4%)
Accelerometer + video TD + audio	54.7% (± 17.2%)	54.8% (± 15.8%)	50.1% (± 12.5%)	65.4% (± 11.6%)	48.1% (± 14.7%)	58.9% (± 12.9%)
All modalities combined	54.6% (± 17.2%)	54.5% (± 12.3%)	50.0% (± 12.4%)	72.0% (± 7.5%)	48.0% (± 14.6%)	61.8% (± 10.7%)

Since computing the precision, recall, and f-measures forces us to discard the posterior probabilities used by the AUC and reduces them to binary classifications, an additional transformation step was necessary here. The binary prediction labels were generated by taking the highest posterior probability for each class. The results are shown in the table below, with standard deviations for all calculated metrics shown in parentheses, respectively.

Authors' Note

We appreciate the research assistance of Navin Raj Prabhu, Marjolein Nanninga, Marissa van der Wel, Sebastian Dorfmeister, Stephanie Tan, Yanxia Zhang, Dina Habib, and Fabeya Kaygun. Our field data collection was supported by Manfred Overmeen, Robert-Jan Sips, Zoltán Szlávik, and Benjamin Timmermans, which is gratefully acknowledged. An abridged version of this paper was presented as a poster at the Annual SIOP Conference 2023 in Boston.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially funded by Netherlands Organization for Scientific Research (NWO) under project number 639.022.606 with associated AspasiaGrant, as well as support by IBM Netherlands.

ORCID iDs

Nale Lehmann-Willenbrock https://orcid.org/0000-0003-3346-5894
Hayley Hung https://orcid.org/0000-0001-9574-5395

References

Acton, B. P., Braun, M. T., & Foti, R. J. (2020). Built for unity: Assessing the impact of team composition on team cohesion trajectories. *Journal of Business and Psychology*, *35*(12), 751-766. https://doi.org/10.1007/s10869-019-09654-7

Alameda-Pineda, X., Ricci, E., & Sebe, N. (2018). *Multimodal behavior analysis in the wild: Advances and challenges*. Academic Press. https://doi.org/10.1016/B978-0-12-814601-9.00007-9.

Allen, J. A., Fisher, C., Chetouani, M., Chiu, M. M., Mehu, M., Gunes, H., & Hung, H. (2017). Comparing social science and computer science workflow processes for studying group interactions. *Small Group Research*, *48*(5), 568-590. https://doi.org/10.1177/1046496417721747

Arrieta, A. B., Rodríguez, N. D., Ser, J. D., Bennetot, A., Tabik, S., & Barbado, A., … & F. Herrera (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bales, R. F. (1950). *Interaction process analysis: A method for the study of small groups*. Addison Wesley.

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423-443. https://doi.org/10.1109/TPAMI.2018.2798607

Banks, G. C., Woznyj, H. M., & Mansfield, C. A. (2021). Where is "behavior" in organizational behavior? A call for a revolution in leadership research and beyond. *The Leadership Quarterly*. Advance online before print. https://doi.org/10.1016/j.leaqua.2021.101581

Barsade, S. G. (2002). The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, *47*(4), 644-675. https://journals.sagepub.com/doi/abs/10.2307/3094912 https://doi.org/10.2307/3094912

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/10.1037/0021-9010.88.6.989

Böck, R. (2021). Affects in groups: A review on automated affect processing and estimation in groups. *IEEE Signal Processing Magazine*, *38*(6), 74-83. https://doi.org/10.1109/MSP.2021.3107811

Bonin, F., De Looze, C., Ghosh, S., Gilmartin, E., Vogel, C., Polychroniou, A., Salamin, H., Vinciarelli, A., & Campbell, N. (2013). Investigating fine temporal dynamics of prosodic and lexical accommodation. In *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, Lyon, France, 25-29 August 2013. http://www.interspeech2013.org/.

Braaten, L. J. (1990). The different patterns of group climate critical incidents in high and low cohesion sessions of group psychotherapy. *International Journal of Group Psychotherapy*, *40*(4), 477-493. https://doi.org/10.1080/00207284.1990.11490623

Brawley, L. R., Carron, A. V., & Widmeyer, W. N. (1987). Assessing the cohesion of teams: Validity of the group environment questionnaire. *Journal of Sport and Exercise Psychology*, *9*(3), 275-294. https://doi.org/10.1123/jsp.9.3.275

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199-231. https://doi.org/10.1214/ss/1009213726

Burgoon, J. K., Magnenat-Thalmann, N., Pantic, M., & Vinciarelli, A. (2017). *Social signal processing*. Cambridge University Press.

Burlingame, G. M., Fuhriman, A., & Johnson, J. E. (2001). Cohesion in group psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, *38*(4), 373. https://doi.org/10.1037/0033-3204.38.4.373

Carmody, P. C., Mateo, J. C., Bowers, D., & McCloskey, M. J. (2017). Linguistic coordination as an unobtrusive, dynamic indicator of rapport, prosocial team processes, and performance in team communication. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 140-144. Sage. https://doi.org/10.1177/1541931213601518

Carron, A. V., Widmeyer, W. N., & Brawley, L. R. (1985). The development of an instrument to assess cohesion in sport teams: The Group Environment Questionnaire. *Journal of Sport and Exercise Psychology*, *7*(3), 244-266. https://doi.org/10.1123/jsp.7.3.244

Casey-Campbell, M., & Martens, M. L. (2009). Sticking it all together: A critical assessment of the group cohesion–performance literature. *International Journal of Management Reviews*, *11*(2), 223-246. https://doi.org/10.1111/j.1468-2370.2008.00239.x

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. https://doi.org/10.1037/a0034142

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079-2107. https://dl.acm.org/doi/abs/10.5555/1756006.1859921

Chartrand, T. L., & Lakin, J. L. (2013). The antecedents and consequences of human behavioral mimicry. *Annual Review of Psychology*, *64*, 285-308. https://doi.org/10.1146/annurev-psych-113011-143754

Chartrand, T. L., & Van Baaren, R. (2009). Human mimicry. *Advances in Experimental Social Psychology*, *41*, 219-274. https://doi.org/10.1016/S0065-2601(08)00405-X

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

Christensen, U., Schmidt, L., Budtz-Jørgensen, E., & Avlund, K. (2006). Group cohesion and social support in exercise classes: Results from a Danish intervention study. *Health Education & Behavior*, *33*(5), 677-689. https://doi.org/10.1177/1090198105277397

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385-390. https://doi.org/10.3758/BRM.41.2.385

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., & Cohen, D. (2012). Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, *3*(3), 349-365. https://doi.org/10.1109/T-AFFC.2012.12

Duffy, K. A., & Chartrand, T. L. (2015). Mimicry: Causes and consequences. *Current Opinion in Behavioral Sciences*, *3*, 112-116. https://doi.org/10.1016/j.cobeha.2015.03.002

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

Eys, M. A., & Carron, A. V. (2001). Role ambiguity, task cohesion, and task self-efficacy. *Small Group Research*, *32*(3), 356-373. https://doi.org/10.1177/104649640103200305

Fischer, T., Hambrick, D. C., Sajons, G. B., & Van Quaquebeke, N. (2020). Beyond the ritualized use of questionnaires: Toward a science of actual behaviors and psychological states. *The Leadership Quarterly*, *31*(4), 101449. https://doi.org/10.1016/S1048-9843(20)30076-X

Gatica-Perez, D., Aran, O., & Jayagopi, D. (2017). Analysis of small groups. In J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, & A. Vinciarelli (Eds.), *Social signal processing* (pp. 349-367). Cambridge University Press.

Gerpott, F. H., Lehmann-Willenbrock, N., & Scheibe, S. (2020). Is work and aging research a science of questionnaires? Moving the field forward by considering perceived versus actual behaviors. *Work, Aging, and Retirement*, *6*(2), 65-70. https://doi.org/10.1093/workar/waaa002

Glass, J. S., & Benshoff, J. M. (2002). Facilitating group cohesion among adolescents through challenge course experiences. *Journal of Experiential Education*, *25*(2), 268-277. https://doi.org/10.1177/105382590202500204

Greatbatch, D., & Clark, T. (2003). Displaying group cohesiveness: Humour and laughter in the public lectures of management gurus. *Human Relations*, *56*(12), 1515-1544. https://doi.org/10.1177/00187267035612004

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

Hemshorn de Sanchez, C. S., Gerpott, F. H., & Lehmann-Willenbrock, N. (2022). A review and future agenda for behavioral research on leader–follower interactions at different temporal scopes. *Journal of Organizational Behavior*, *43*(2), 342-368. https://doi.org/10.1002/job.2583

Hill, N. S., Offermann, L. R., & Thomas, K. (2019). Mitigating the detrimental impact of maximum negative affect on team cohesion and performance through face-to-face communication. *Group & Organization Management*, *44*(1), 211-238. https://doi.org/10.1177/1059601118776835

Hoehl, S., Fairhurst, M., & Schirmer, A. (2021). Interactional synchrony: Signals, mechanisms and benefits. *Social Cognitive and Affective Neuroscience*, *16*(1-2), 5-18. https://doi.org/10.1093/scan/nsaa024

Holmes, J., & Marra, M. (2006). Humor and leadership style. *Humor*, *19*(2), 119-138. https://doi.org/10.1515/HUMOR.2006.006

Hoogeboom, M. A., & Wilderom, C. P. (2020). A complex adaptive systems approach to real-life team interaction patterns, task context, information sharing, and effectiveness. *Group & Organization Management*, *45*(1), 3-42. https://doi.org/10.1177/1059601119854927

Hospedales, T. M., Antoniou, A., Micaelli, P., & Storkey, A. J. (2022). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*, 5149-5169. https://doi.org/10.1109/TPAMI.2021.3079209

Hüffmeier, J., & Hertel, G. (2011). Many cheers make light the work: How social support triggers process gains in teams. *Journal of Managerial Psychology*, *26*(3), 185-204. https://doi.org/10.1108/02683941111112631

Hung, H., & Gatica-Perez, D. (2010). Estimating cohesion in small groups using audio–visual nonverbal behavior. *IEEE Transactions on Multimedia*, *12*(6), 563-575. https://doi.org/10.1109/TMM.2010.2055233

Hung, H., Murray, G., Varni, G., Lehmann-Willenbrock, N., Gerpott, F. H., & Oertel, C. (2020, October). Workshop on interdisciplinary insights into group and team dynamics. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 876-877). https://doi.org/10.1145/3382507.3419748.

Jackson, J. C., Jong, J., Bilkey, D., Whitehouse, H., Zollmann, S., McNaughton, C., & Halberstadt, J. (2018). Synchrony and physiological arousal increase cohesion and cooperation in large naturalistic groups. *Scientific Reports*, *8*(1), 1-8. https://doi.org/10.1038/s41598-017-18023-4

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.). Springer.

Kapcak, Ö, Vargas-Quiros, J., & Hung, H. (2019, September). Estimating romantic, social, and sexual attraction by quantifying bodily coordination using wearable sensors. In 2019 *8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (pp. 154-160). IEEE. https://doi.org/10.1109/ACIIW.2019.8925137.

Kauffeld, S., & Lehmann-Willenbrock, N. (2012). Meetings matter: Effects of team meeting communication on team and organizational success. *Small Group Research*, *43*(2), 128-156. https://doi.org/10.1177/1046496411429599

Kauffeld, S., Lehmann-Willenbrock, N., & Meinecke, A. L. (2018). The advanced interaction analysis for teams (act4teams) coding scheme. In E. Brauner, M. Boos, & M. Kolbe (Eds.), *The Cambridge handbook of group interaction analysis* (pp. 422-431). Cambridge University Press.

Keyton, J. (2018). Interaction analysis: An introduction. In E. Brauner, M. Boos, & M. Kolbe (Eds.), *The Cambridge handbook of group interaction analysis* (pp. 3-19). Cambridge University Press.

Keyton, J., & Heylen, D. K. (2017). Pushing interdisciplinary in the study of groups and teams. *Small Group Research*, *48*(5), 621-630. https://doi.org/10.1177/1046496417732528

Khalid, S., Khalil, T., & Nasreen, S. (2014, August). A survey of feature selection and feature extraction techniques in machine learning. 2014 *Science and Information Conference*, 2014, pp. 372-378. https://doi.org/10.1109/SAI.2014.6918213.

Klonek, F. E., Gerpott, F. H., Lehmann-Willenbrock, N., & Parker, S. (2019). Time to go wild: How to conceptualize and measure process dynamics in real teams with high resolution. *Organizational Psychology Review*, *9*(4), 245-275. https://doi.org/10.1177/2041386619886674

Kolbe, M., & Boos, M. (2019). Laborious but elaborate: The benefits of really studying team dynamics. *Frontiers in Psychology*, *10*, 1478. https://doi.org/10.3389/fpsyg.2019.01478

Kozlowski, S. W. (2015). Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations. *Organizational Psychology Review*, *5*(4), 270-299. https://doi.org/10.1177/2041386614533586

Kozlowski, S. W., & Chao, G. T. (2012). The dynamics of emergence: Cognition and cohesion in work teams. *Managerial and Decision Economics*, *33*(5-6), 335-354. https://doi.org/10.1002/mde.2552

Kozlowski, S. W., & Chao, G. T. (2018). Unpacking team process dynamics and emergent phenomena: Challenges, conceptual advances, and innovative methods. *American Psychologist*, *73*(4), 576. https://doi.org/10.1002/mde.2552

Lakin, J. L. (2013). Behavioral mimicry and interpersonal synchrony. In J. A. Hall, & M. L. Knapp (Eds.), *Handbooks of communication science. Nonverbal communication* (pp. 539-575). De Gruyter Mouton. https://doi.org/10.1515/9783110238150.539

LeBaron, C., Jarzabkowski, P., Pratt, M. G., & Fetzer, G. (2018). An introduction to video methods in organizational research. *Organizational Research Methods*, *21*(2), 239-260. https://doi.org/10.1177/1094428117745649

Lehmann-Willenbrock, N., & Allen, J. A. (2014). How fun are your meetings? Investigating the relationship between humor patterns in team interactions and team performance. *Journal of Applied Psychology*, *99*(6), 1278-1287. https://doi.org/10.1037/a0038083

Lehmann-Willenbrock, N., & Allen, J. A. (2018). Modeling temporal interaction dynamics in organizational settings. *Journal of Business and Psychology*, *33*(3), 325-344. https://doi.org/10.1007/s10869-017-9506-9

Lehmann-Willenbrock, N., Allen, J. A., & Kauffeld, S. (2013). A sequential analysis of procedural meeting communication: How teams facilitate their meetings. *Journal of Applied Communication Research*, *41*(4), 365-388. https://doi.org/10.1080/00909882.2013.844847

Lehmann-Willenbrock, N., & Chiu, M. M. (2018). Igniting and resolving content disagreements during team interactions: A statistical discourse analysis of team dynamics at work. *Journal of Organizational Behavior*, *39*(9), 1142-1162. https://doi.org/10.1002/job.2256

Lehmann-Willenbrock, N., Chiu, M. M., Lei, Z., & Kauffeld, S. (2017a). Understanding positivity within dynamic team interactions: A statistical discourse analysis. *Group & Organization Management*, *42*(1), 39-78. https://doi.org/10.1177/1059601116628720

Lehmann-Willenbrock, N., Hung, H., & Keyton, J. (2017b). New frontiers in analyzing dynamic group interactions: Bridging social and computer science. *Small Group Research*, *48*(5), 519-531. https://doi.org/10.1177/1046496417718941

Lichtenauer, J., Shen, J., Valstar, M., & Pantic, M. (2011). Cost-effective solution to synchronised audio–visual data capture using multiple sensors. *Image and Vision Computing*, *29*(10), 666-680. https://doi.org/10.1016/j.imavis.2011.07.004

Lott, A. J., & Lott, B. E. (1965). Group cohesiveness as interpersonal attraction: A review of relationships with antecedent and consequent variables. *Psychological Bulletin*, *64*(4), 259-309. https://doi.org/10.1037/h0022386

Mast, M., Gatica-Perez, D., Frauendorfer, D., Nguyen, L., & Choudhury, T. (2015). Social sensing for psychology: Automated interpersonal behavior assessment. *Current Directions in Psychological Science*, *24*, 154-160. https://doi.org/10.1177/0963721414560811

Mayo, O., & Gordon, I. (2020). In and out of synchrony—Behavioral and physiological dynamics of dyadic interpersonal coordination. *Psychophysiology*, *57*(6), e13574. https://doi.org/10.1111/psyp.13574

Moreland, R. L. (2010). Are dyads really groups? *Small Group Research*, *41*(2), 251-267. https://doi.org/10.1177/1046496409358618

Müller, J., Fàbregues, S., Guenther, E. A., & Romano, M. J. (2019). Using sensors in organizational research—clarifying rationales and validation challenges for mixed methods. *Frontiers in Psychology*, *10*, 1188. https://doi.org/10.3389/fpsyg.2019.01188

Muller, P. M., & Bulling, A. (2019). Emergent leadership detection across datasets. In *ICMI '19: 2019 International Conference on Multimodal Interaction* (274-278). https://doi.org/10.1145/3340555.3353721

Murray, G., & Oertel, C. (2018, October). Predicting group performance in task-based interaction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (pp. 14-20).

Nanninga, M. C., Zhang, Y., Lehmann-Willenbrock, N., Szlávik, Z., & Hung, H. (2017, November). Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (pp. 206-215). https://doi.org/10.1145/3136755.3136811

Pantic, M., Cowie, R., D'Errico, F., Heylen, D., Mehu, M., Pelachaud, C., Poggi, I., Schroeder, M., & Vinciarelli, A. (2011). Social signal processing: The research agenda. In T. B. Moeslund, A. Hilton, V. Krüger, & L. Sigal (Eds.), *Visual analysis of humans* (pp. 511-538). Springer.

Pentland, A., & Heibeck, T. (2008). *Honest signals*. *MIT press*.

Poppe, R. (2011). Automatic analysis of bilidy social signals. In J. Burgoon, N. Magnenat-Thalmann, M. Pantic, & A. Vinciarelli (Eds.), *Social signal processing* (pp. 155-167). Cambridge University Press. https://doi.org/10.1017/9781316676202.001

Raman, C., Nonnemaker, A., Villegas-Morcillo, A., Hung, H., & Loog, M. (2022). Why did this model forecast this future? Closed-form temporal saliency towards causal explanations of probabilistic forecasts. *arXiv preprint arXiv:2206.00679*. Available at https://arxiv.org/pdf/2206.00679.

Raman, C., Tan, S., & Hung, H. (2020). A modular approach for synchronized wireless multimodal multisensor data acquisition in highly dynamic social settings. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 3586-3594). https://doi.org/10.1145/3394171.3413697

Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., & Herrera, F. (2017). A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, *239*, 39-57. https://doi.org/10.1016/j.neucom.2017.01.078

Rönkkö, M., Aalto, E., Tenhunen, H., & Aguirre-Urreta, M. I. (2022). Eight simple guidelines for improved understanding of transformations and nonlinear effects. *Organizational Research Methods*, *25*(1), 48-87. https://doi.org/10.1177/1094428121991907

Rosopa, P. J. (2022). Machine learning and the science of teams. In B. Murray, J. Dulebohn, & D. Stone (Eds.), *Managing team centricity in modern organizations* (pp. 53-76). Information Age.

Rudovic, O., Nicolaou, M. A., & Pavlovic, V. (2017). Machine learning methods for social signal processing. In J. Burgoon, N. Magnenat-Thalmann, M. Pantic, & A. Vinciarelli (Eds.), *Social signal processing* (pp. 234-254). Cambridge University Press.

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

Santoro, J. M., Dixon, A. J., Chang, C. H., & Kozlowski, S. W. (2015). Measuring and monitoring the dynamics of team cohesion: Methods, emerging tools, and advanced technologies. In E. Salas, W. B. Vessey, & A. X. Estrada (Eds.), *Team cohesion: Advances in psychological theory, methods and practice (Research on managing groups and teams*) (Vol. 17, pp. 115-145). Emerald. https://doi.org/10.1108/S1534-085620150000017006

Scogin, F. R., & Pollio, H. R. (1980). Targeting and the humorous episode in group process. *Human Relations*, *33*(11), 831-852. https://doi.org/10.1177/001872678003301105

Shields, D. L. L., Gardner, D. E., Bredemeier, B. J. L., & Bostrom, A. (1995). Leadership, cohesion, and team norms regarding cheating and aggression. *Sociology of Sport Journal*, *12*(3), 324-336. https://doi.org/10.1123/ssj.12.3.324

Shuffler, M. L., & Cronin, M. A. (2019). The challenges of working with "real" teams: Challenges, needs, and opportunities. *Organizational Psychology Review*, *9*(4), 211-218. https://doi.org/10.1177/2041386620901884

Solanki, V., Vinciarelli, A., Stuart-Smith, J., & Smith, R. (2016). When the game gets difficult, then it is time for mimicry. In A. Esposito, M. Faundez-Zanuy, A. M. Esposito, G. Cordasco, T. Drugman, J. Solé-Casals, & F. C. Morabito (Eds.), *Recent advances in nonlinear speech processing* (pp. 247-254). Springer.

Stel, M., & Vonk, R. (2010). Mimicry in social interaction: Benefits for mimickers, mimickees, and their interaction. *British Journal of Psychology*, *101*(2), 311-323. https://doi.org/10.1348/000712609X465424

Terrion, J. L., & Ashforth, B. E. (2002). From 'I'to 'we': The role of putdown humor and identity in the development of a temporary group. *Human Relations*, *55*(1), 55-88. https://doi.org/10.1177/0018726702055001606

Tschacher, W., Rees, G. M., & Ramseyer, F. (2014). Nonverbal synchrony and affect in dyadic interactions. *Frontiers in Psychology*, *5*, 1323. https://doi.org/10.3389/fpsyg.2014.01323

Uitdewilligen, S., Rico, R., & Waller, M. J. (2018). Fluid and stable: Dynamics of team action patterns and adaptive outcomes. *Journal of Organizational Behavior*, *39*(9), 1113-1128. https://doi.org/10.1002/job.2267

Urban, C. J., & Gates, K. M. (2021). Deep learning: A primer for psychologists. *Psychological Methods*, *26*(6), 743-773. https://doi.org/10.1037/met0000374

Vanhove, A. J., & Herian, M. N. (2015). Team cohesion and individual well-being: A conceptual analysis and relational framework. In E. Salas, W. B. Vessey, & A. X. Estrada (Eds.), *Team cohesion: Advances in psychological theory, methods and practice (research on managing groups and teams*. (Vol. 17, pp. 53-82). Emerald. https://doi.org/10.1108/S1534-085620150000017004.

Vanschoren, J. (2019). Meta-learning. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated machine learning. The Springer series on challenges in machine learning* (pp. 35-61). Springer.

Vinciarelli, A. (2017). Introduction: Social signal processing. In J. Burgoon, N. Magnenat-Thalmann, M. Pantic, & A. Vinciarelli (Eds.), *Social signal processing* (pp. 1-8). Cambridge University Press. https://doi.org/10.1017/9781316676202.001

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, *27*(12), 1743-1759. https://doi.org/10.1016/j.imavis.2008.11.007

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., & Schroeder, M. (2011). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, *3*(1), 69-87. https://doi.org/10.1109/T-AFFC.2011.27

Waller, M. J., & Kaplan, S. A. (2018). Systematic behavioral observation for emergent team phenomena: Key considerations for quantitative video-based approaches. *Organizational Research Methods*, *21*(2), 500-515. https://doi.org/10.1177/1094428116647785

Wang, H., Kläser, A., Schmid, C., & Liu, C. L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, *103*(1), 60-79. https://doi.org/10.1007/s11263-012-0594-8

Weiss, G. M. (2013). Foundations of imbalanced learning. In H. He, & Y. Ma (Eds.), *Imbalanced learning: Foundations, algorithms, and applications* (pp. 13-42). Wiley.

West, B. J., Patera, J. L., & Carsten, M. K. (2009). Team level positivity: Investigating positive psychological capacities and team level outcomes. *Journal of Organizational Behavior*, *30*(2), 249-267. https://doi.org/10.1002/job.593

Whitton, S. M., & Fletcher, R. B. (2014). The group environment questionnaire: A multilevel confirmatory factor analysis. *Small Group Research*, *45*(1), 68-88. https://doi.org/10.1177/1046496413511121

Wilson, S., & Gos, C. (2019). Perceiving social cohesion: Movement synchrony and task demands both matter. *Perception*, *48*(4), 316-329. https://doi.org/10.1177/0301006619837878

Witten, I. H., Frank, E., & Hall, M. A. (2005). *Practical machine learning tools and techniques* (3rd ed). Elsevier.

Zaccaro, S. J., Gualtieri, J., & Minionis, D. (1995). Task cohesion as a facilitator of team decision making under temporal urgency. *Military Psychology*, *7*(2), 77-93. https://doi.org/10.1207/s15327876mp0702_3

Zhang, Y., Olenick, J., Chang, C. H., Kozlowski, S. W., & Hung, H. (2018). Teamsense: Assessing personal affect and group cohesion in small teams through dyadic interaction and behavior analysis with wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1-22. https://doi.org/10.1145/3264960

Zhu, Q., Yeh, M. C., Cheng, K. T., & Avidan, S. (2006, June). Fast human detection using a cascade of histograms of oriented gradients. In 2006 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 1491-1498. IEEE.

Author Biographies

Nale Lehmann-Willenbrock is professor and department chair of Industrial/Organizational Psychology,director of the Center for Better Work, and vice dean of the Faculty of Psychology and Movement Science at the University of Hamburg, Germany. She investigates team processes, leader–follower dynamics, and organizational meetings through a behavioral lens, using quantitative interaction analysis and various tools to detect systematic patterns of social interaction.

Hayley Hung is an associate professor of Computer Science and leads the Perceptive Computing Lab at Delft University of Technology, The Netherlands. Her research focuses on how automated systems can become aware of people as social beings. She devises models to automatically interpret face-to-face human social behavior using cameras, microphones, and wearable sensors.