Measuring Together, Thinking Apart: Interrogating Alignment In Group Informatics Systems

Sharon Ferguson
University of Waterloo
Waterloo, ON, Canada
sharon.ferguson@uwaterloo.ca

Zayn Dhillon
University of Waterloo
Waterloo, ON, Canada
zdhillon@uwaterloo.ca

Avelyn Wong
University of Toronto
Toronto, ON, Canada
avelyn.wong@mail.utoronto.ca

Georgia Van de Zande
Olin College of Engineering
Needham, MA, USA
gvandezande@olin.edu

Alison Olechowski
University of Toronto
Toronto, ON, Canada
a.olechowski@utoronto.ca

ABSTRACT

Modern knowledge work is increasingly collaborative and fragmented across space and time. The traces from virtual collaboration platforms could be used to support effective collaboration by measuring group-level teamwork characteristics. We call these proposed systems group informatics, which analyze and allow teams to reflect on collaboration dynamics. When systems consolidate and present data from a variety of stakeholders, team members are likely to disagree on what successful collaboration looks like. In this work, we present an exploratory study of (mis)alignment in group informatics systems, focusing on human-human (H-H) alignment, human-algorithm (H-A) alignment, and algorithm-algorithm (A-A) alignment. Using a parallel mixed-methods study across two populations, we collect qualitative perceptions of alignment using a functional technology probe that measures shared understanding through four time-series graphs derived from team Slack messages, in an in-situ study. We complement this with a larger-scale quantitative investigation. Across two populations, we found little evidence of alignment of any type. We find that these teamwork constructs are complex—algorithmic measures capture different facets of the construct, H-H alignment presents a barrier to H-A alignment, and visual presentation of the algorithms influences alignment. We further find that alignment is influenced by extreme or unexpected events. We discuss implications for the design of holistic, co-constructed measures in group informatics systems, the appropriate granularity of algorithmic feedback, and whether alignment is achievable or necessary.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

GI '26, June 9-12, 2026, Waterloo, ON

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

KEYWORDS

Group Informatics, Collaboration, Alignment, Digital Trace Data, Shared Understanding

ACM Reference Format:
Sharon Ferguson, Zayn Dhillon, Avelyn Wong, Georgia Van de Zande, and Alison Olechowski. 2026. Measuring Together, Thinking Apart: Interrogating Alignment In Group Informatics Systems. In *Proceedings of Graphics Interface (GI '26).* ACM, New York, NY, USA, 20 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Modern knowledge work is becoming increasingly collaborative, with workers spending up to 80% of their time collaborating [105]. This collaboration is often spread across time, space [85], communication modalities (verbal, text, video) and platforms (chat, shared documents, video calls) [44, 113]. This fragmentation can make it challenging for teams to capture a complete picture of their collaboration [112], and for managers to provide collaboration feedback. In fact, team collaboration dynamics are largely invisible [33], and are often not recognized until it is too late [31].

With more and more work happening on virtual platforms [2, 44], there is the potential for these systems to sense or provide feedback on collaboration dynamics. We refer to these future technologies as *group informatics systems*, defined as systems that present metrics and summaries of important collaboration characteristics, which are reported for group reflection. Commercially, virtual collaboration platforms are already implementing simple metrics for collaboration feedback, including Zoom's smart AI assistant[1], which can provide real-time feedback on talking patterns; ReadAI's meeting plug in[2] which comments on sentiment and engagement; and WebEx's collaboration insights[3] focused on collaboration time and scheduling. HCI researchers are starting to explore this area, with a design fiction highlighting the challenges of these systems [73], and surveys exploring which types of collaborative metrics are of interest [79]. Similar to these tools, we develop a functional technology probe, SharUn, to visualize collaboration over time (see Figure 4).

[1] https://www.zoom.com/en/blog/zoom-ai-companion/
[2] https://support.read.ai/hc/en-us/articles/4406653674003-About-Sentiment-Engagement-and-the-Read-Score
[3] https://www.webex.com/collaboration-insights.html

Many collaborative measures are, by definition, emergent and collective phenomena [73]. Take psychological safety [41] as an example, where researchers verify internal consistency within their dataset before claiming the measure as a valid team metric [26]. This suggests that some measures of successful collaboration will require agreement across an entire team, which may be challenging when systems involve multiple stakeholders, who may each have different understandings of what successful collaboration dynamics look like, or how success should be measured [103]. Lallemand et al. [73] cautions that when dealing with group-level metrics, we need to question the assumptions that are baked into the "optimal" levels of each metric. Misalignment among team members on what is measured or how measurements are interpreted can lead to mistrust in the system, if workers do not see the output as useful or relevant [56], or more harmful consequences if leaders interpret system outputs differently from team members [81, 87]. As group informatics is a new area of research, we know little about what forms of alignment these systems require, how this alignment manifests, and to what extent alignment is feasible or required.

In this work, we focus on three forms of alignment in group informatics systems (see Figure 1). **Human-Human (H-H)** alignment refers to differences in how successful teamwork is perceived across members of the team, **Human-Algorithm (H-A)** alignment refers to differences in how automated systems measure success and how humans perceive it, and **Algorithm-Algorithm (A-A)** alignment refers to the difference between various automated measures of successful teamwork. Past work has focused primarily on Human-Algorithm alignment, such as how humans and AI measure emotion in the workplace [68]. Designing and implementing group informatics systems in ways that reflect the realities of each form of (mis)alignment will encourage their use as flexible reflection tools, and help prevent feelings of mistrust toward the tool and team members.

To investigate alignment, we focus on one element of successful teamwork as a case study: shared understanding. We chose shared understanding because it is a team-level construct important for successful teams [75] and has several proposed text-based measurements [35, 62, 78]. Shared understanding is often described as the similarity in the way individuals perceive, encode, store, and retrieve information, required to successfully work together [74] and adapt to changing conditions [42]. Using this focus, we investigate the research question: **How does (mis)alignment manifest across human-human, human-algorithm, and algorithm-algorithm dimensions in a shared understanding group informatics systems, and what influences it?** Human-Algorithm (H-A) Alignment, and Human-Human (H-H) Alignment.

We conduct an exploratory, mixed-methods study with two populations: a professional, in-situ population that captures in-depth, qualitative reflections; and a longitudinal, larger-scale study of students for quantitative estimates of alignment. We develop a functional technology probe, SharUn, which applies existing shared understanding algorithms to public channel Slack messages and visualizes this data to users via four time-series graphs, used in the in-situ study.

We found little evidence of alignment across the board. Identifying two main themes, construct complexity and the influence of extreme events, we focus on four key findings. Quantitatively, algorithmic measures of shared understanding show little Algorithm-Algorithm alignment, yet in-situ participants described the measures as complementary, suggesting that A-A misalignment reflects construct multidimensionality. The user study participants noted varying signals of shared understanding that were not captured by any algorithmic measures, such as agreement, demonstrating that Human-Human alignment is a barrier to Human-Algorithm alignment. The user study also revealed that the presentation of the algorithmic output systematically shapes user interpretation and, thus, alignment. Our qualitative study identified that H-H and H-A alignment was influenced by extremes or key events. Spikes in algorithmic measures corresponded with key events, such as software outages or brainstorming sessions, during which they expected shared understanding to be very high or very low. Yet the algorithms also identified day-to-day variation that team members did not perceive.

From these findings, we begin a discussion of the importance of carefully designed holistic measures that capture the multiple dimensions of successful teamwork, and of how we can include more voices in the design of these systems. We discuss whether day-to-day variation is important for users to be aware of, and ask whether alignment is achievable or even necessary. We contribute an early characterization of alignment in group informatics systems across three dimensions, surfacing key design questions to consider as we continue to investigate the applicability of automated teamwork systems.

2 BACKGROUND

We review literature on group informatics, alignment processes in teams, and teamwork dynamics with a focus on shared understanding.

2.1 Informatics Systems

HCI researchers have begun to consider systems for group feedback and reflection, including questions about what should be calculated [79] and what consequences could arise [73]. Lushnikova et al. [79] looked into what aspects of collaboration people would be interested in learning about, finding that participants are interested in tracking metrics that are task-oriented, such as number of emails read; relation-oriented, such as discrimination or impact of hierarchy; individual-oriented, such as emotions and belonging; time-oriented, such as scheduling; and space-oriented, such as the impact of one's office door being opened or closed. Notably, most of these metrics were *about* collaboration yet were *individually-measured*. Existing systems also provide individual measures of collaboration. Visualizations have been developed to provide real-time participation equality feedback to each participant in a team discussion [11, 76], and individualized feedback based on linguistic characteristics [76]. In these types of systems, individuals preferred being able to see others' data in order to understand and contextualize their own [40, 47]. In these examples, presenting collaboration metrics to groups may simply mean aggregating individual measures.

Group informatics systems can also measure *group-level constructs* that don't exist at an individual level, such as conflict or

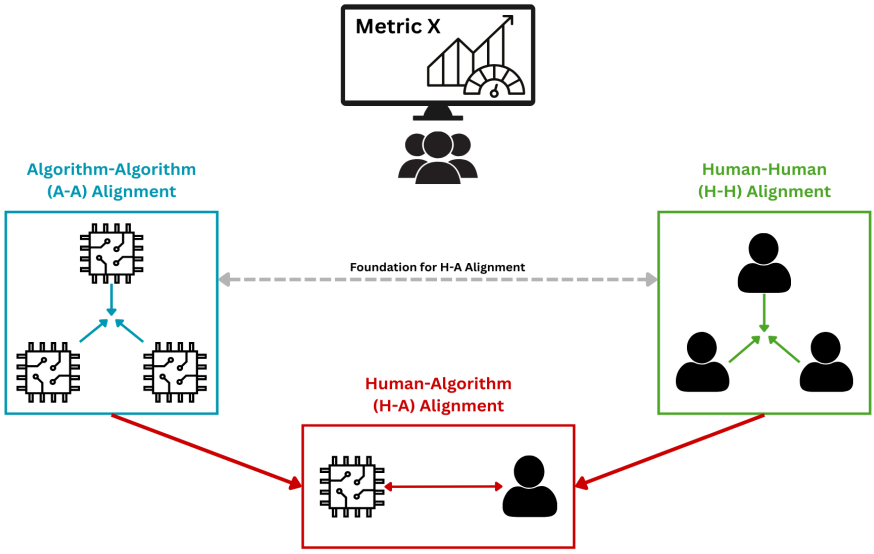

Figure 1: Types of alignment studied in this paper: algorithm-algorithm (A-A) alignment, human-algorithm (H-A) alignment, and human-human (H-H) alignment

psychological safety [41]. For example, Almutairi et al. [3] developed a chat application that provided teams with feedback on their collaboration by measuring group-level attributes such as linguistic alignment and topic coherence. Further, Chen et al. [22] developed passive and active versions of a prototype system to provide feedback to teams during video meetings, ensuring that they stay on track with the meeting's goals. Providing data-based feedback to team members has been shown to improve shared understanding and creativity [88]. A post-meeting dashboard containing participation, sentiment, consensus, questions, and turn-taking improved participants' awareness of meeting dynamics [97] and balance in conversation [98].

These types of systems also come with a number of challenges, ethical considerations, and open research directions. One of the most commonly noted concerns for systems used in a workplace setting is that their metrics may be used for unintended purposes, such as to fire or prevent career progression [81], or for general evaluation and performance management [87]. Workers are also concerned about how different data types will be shared with colleagues [103]. Implementing productivity-measuring systems also has implications for trust between employees and the organization [81]. Merely being aware that data is being collected can trigger the Hawthorne Effect, where workers act differently because they know they are being monitored [32], and monitoring systems may foster workplace competition [69, 103]. Lallemand et al. [73]'s design fiction explores a number of concerns that may materialize in the future—including measuring group-level attributes that are more than the sum of individual measures, carefully considering the assumptions baked into these metrics, aligning the measures with the team's understanding of collaboration, and encouraging feedback to be reviewed and interpreted as a team. The authors also question to what extent complex social processes can be quantified without reducing them, while protecting existing intra-team trust.

Another commonly noted challenge is alignment [103]. Suh et al. [103] studied the implementation of a mental wellbeing sensing technology in the workplace, and identified five categories of misalignments, including misalignments in boundaries between work and non-work, data ownership beliefs, values and motivations for personal informatics (wellbeing vs. productivity), optimal metric definitions, and levels of power. In this work, we focus on alignment in the *metric definitions*, where members may have different ideas of what "optimal" entails for a specific metric [107], and may not agree with automated measures [68] (Figure 1). In a study of emotion recognition in the workplace, Kaur et al. [68] found that the automated measures did not align with humans' perceptions of their emotions. They also note that people are biased in their own reporting, and complete alignment may not always be the goal. Instead, the authors suggest we evaluate whether automation captures useful information.

In this work, we narrow in on the challenge of alignment in group informatics systems, studying it across three levels. To understand the extent to which we might expect alignment and how teams achieve it, we review research on collective sensemaking.

2.2 Collective Sensemaking and Alignment Processes

For work groups to function, members must communicate in ways that enable coordinated action, including the collective ability to understand different facets of the problem, share their insights openly, and build commitment to the final decision [19]. Collaborative sensemaking emphasizes a group's ability to form, organize, and reach consensus when faced with novel or ambiguous problems [39]. In practice, sensemaking is iterative and distributed, requiring teams to continuously update their interpretations as situations evolve [65]. A central mechanism supporting collaborative sensemaking is alignment. Successful communication depends on aligning representations so that team members develop similar situation models, allowing coordinated action without explicit negotiation [92]. Through interaction, teams gradually establish common ground that supports shared interpretations.

As data becomes central to organizational work, teams increasingly engage in sensemaking around quantitative representations. Visualizations support sensemaking by helping people collect, organize, and analyze information to generate knowledge and inform action [57]. However, sensemaking with data is often a social activity, involving discussion, comparison of interpretations, and consensus building [57]. Collaborative visualization systems explicitly support this joint interpretation by enabling multiple people to interact with shared data representations [8, 64]. This involves ongoing negotiation over what information counts, how it should be interpreted, and how it applies to different roles within the group [39]. Effective group informatics systems may need to support the social processes through which teams align interpretations. By enabling interaction, discussion, and comparison of perspectives, such systems can help teams develop insights that would not emerge from individual analysis alone [64, 72].

While collective sensemaking is ongoing, it becomes particularly explicit and intensive when expectations are disrupted. Sensemaking is triggered by events that are novel, ambiguous, confusing, or violate expectations [80]. Weick et al. [110] describes these as moments when an expectation of continuity is breached, including discrepancy, breakdown, surprise, or interruption. In knowledge work settings, such disruptions can include wasted time, unfair practices, improper work, and process disruptions [37]. This suggests that alignment processes may differ between routine and uncommon situations: in routine situations, teams may need only minimal alignment to move forward, whereas extreme events may force them to address misalignment directly.

Collaborative sensemaking is required for group informatics systems as team dynamics, and thus feedback and measures related to these dynamics, are nuanced and multidimensional [71]. Next, we review the multidimensional and emergent nature of teamwork, and provide a background on our case study construct of interest, shared understanding.

2.3 Multidimensional Nature of Teamwork

Measuring successful teamwork inherently requires a multidimensional approach [1, 71]. Recent HCI work that aims to characterize or encourage successful teamwork incorporates a number of complementary measures [3, 18]. This multidimensional nature has created measurement challenges in team science research [1].

Research on teams emphasizes that group-level phenomena are emergent rather than aggregative. Teams rely on both individual and mutual accountability, producing outcomes that exceed the sum of individual contributions [67]. Similarly, research on team cognition shows that shared understanding does not arise from identical mental states, but from compatible and congruent knowledge structures that support coordinated action [96]. As a result, group-level constructs such as understanding, coordination, and effectiveness must be studied as products of interaction rather than as averages of individual perspectives.

2.3.1 Shared Understanding. Shared understanding, often referred to as shared/team mental models [13, 75] or common ground [25], is commonly defined as "the degree of cognitive overlap and commonality in beliefs, expectations, and perceptions about goals, processes, tasks, members' knowledge, skills, and abilities" [59]. Shared understanding in teams has a number of benefits, including improved anticipation of other team members [59], fewer duplicated efforts, efficient use of resources, increased satisfaction, increased creativity [99], and higher performance [83, 84]. However, a perfect shared understanding may result in groupthink [29, 75].

Given the benefits of shared understanding, researchers have studied team-level factors, finding teams with similar backgrounds and shared experiences are better off [59]; process factors, finding that shared understanding is developed over time and interactions [13]; and tool factors, finding benefits of open communication channels [38], video calling [108], collaborative documentation [13], informal communication [89], persistent chats [49], and personal profiles [111]. Teams collaborating via computer-mediated communication needed to clarify misunderstandings more often [4, 59], and reached a shared understanding later in their lifecycle [82]. Contributing factors here include the absence of body language and tone, making it more challenging to recognize when a team member is confused and limits the ability to adjust communication in real-time [4].

As with any cognitive process, shared understanding is challenging to measure. Researchers commonly use structured surveys that capture how different elements of a shared task are related [6, 10, 21, 27, 66]. Researchers also use survey perceptions of shared understanding [27, 28], or content analysis of transcripts [30, 115]. There are calls for more global [84] and continuous [29] measures of shared understanding, such as those used here. Lix et al. [78] argues that the natural language that teams use to communicate can represent group-level shared cognition, whereas surveys measure only individual, episodic perceptions of shared cognition. Three such natural language processing methods are Language Style Matching (LSM), Latent Semantic Analysis (LSA), and embedding spaces.

Language Style Matching (LSM) compares the frequency with which team members use categories of function words [62] and describes how things were said rather than what was said. Language styles are strong indicators of having common ground [7, 61], and a shared perspective [90], and have previously been used to evaluate teamwork [3]. Latent Semantic Analysis (LSA) is another

common measure [35, 36, 58, 114] that captures shared understanding by characterizing topic and voice similarity in documents [58] or transcripts [34]. Babcock et al. [7] found that LSA measures increase with highly involved interactions, while LSM increases with strong emotions. We include two LSA-derived metrics in this work: knowledge convergence and semantic coherence [35]. The last commonly used measure utilizes embedding spaces, which are models that transform words and sentences into vectors that retain the meaning. We compare embedding space similarity to LSM and LSA, as embedding space models allow for a larger corpus and better capture underlying word meanings [23].

In summary, past work suggests that alignment will be a challenge when deriving automated measures of complex constructs [68, 103], and this challenge intensifies when alignment is required not just across a single human and a single system, but between groups of humans and a system. Focusing on shared understanding as an example of a complex teamwork construct, we unpack the challenge of alignment across Human-Algorithm, Human-Human, and Algorithm-Algorithm alignment.

3 METHODS

We conduct a mixed-method, exploratory study across two participant groups to present a multi-context investigation of alignment. In our first study, we qualitatively examined alignment by conducting an in-situ study with seven participants across two academic and two industry teams as they interacted with a technology probe. In the second study, we quantitatively analyze Slack messages and survey responses from students in a mechanical engineering capstone course using the same algorithms as in the in-situ study, applied post-hoc. An overview of the methods is shown in Figure 2.

This two-population study allowed us to examine alignment from distinct but complementary vantage points, as recommended for the complexity of shared understanding [5]. The in-situ qualitative study placed participants in a realistic collaborative context so we could reflect their data as accurately as possible, while the quantitative study allowed us to examine algorithmic alignment on a larger scale. While the qualitative study provides insight into why misalignment occurs and participant perceptions, the quantitative study provides a naturalistic baseline not influenced by the reflection prompted by the technology probe. Additionally, the quantitative study captures alignment longitudinally (three-month timeline), whereas the qualitative study captures post-hoc reflection of a two-week period—previous research suggests that alignment and shared mental models change over time [13, 25]. Further, the qualitative study enabled in-depth reflection, which is not possible with traditional survey-based measures of shared understanding, and the quantitative study enabled comparison with these traditional measures at a larger scale. Both studies address different aspects of the alignment question: the qualitative study focuses on Human-Human and Human-Algorithm alignment through participant reflection, while the quantitative study enables comparison across H-A and A-A alignment using validated survey measures. By assessing these two populations, we can develop hypotheses about the mechanisms that influence perceptions of alignment across technology, time, and scale.

3.1 Algorithms

As shown in Section 2.3.1, shared understanding can be signalled in multiple ways, such as mimicking one's language style [63], converging in vocabulary (using the same terms in consistent ways) [7], or using different terms when the underlying meaning remains the same [78]. As such, we used a variety of algorithms to capture multifaceted perceptions of shared understanding. The algorithms described here are applied to both the qualitative in-situ study and the quantitative post-hoc study. Similar to Lix et al. [78], our measures of shared understanding encompass the full range of meanings communicated between group members and are not specific to a single element, such as a task or team.

Data Preprocessing: We began by removing all links and non-dictionary words using the enchant package [70]. We then aggregated all messages sent by each individual in a specific channel on a specific day, as the algorithms are calculated per day, and many compare individuals. After this processing, the input to the following algorithms is cleaned Slack messages, aggregated by channel-day-user. We measure shared understanding using four different algorithms previously proposed in past work, which we adapted to the Slack short-messaging context where appropriate. The result of each algorithm, from the same channel (of the authors' Slack) and time period, are shown in Figure 3. Additional details on all algorithms are shown in Appendix A.

Language Style Matching (LSM): We implemented the LSM algorithm as defined by Ireland and Pennebaker [62], measuring the frequency of linguistic categories (pronouns, articles, prepositions, negations, verbs, conjugations) through the Linguistic Inquiry and Word Count 2022 dictionary [15] accessed through the Receptiviti API [95]. We compared these measures across all potential pairs of team members.

Knowledge Convergence: Knowledge convergence was originally defined by Dong [35] in his study of engineering designer communication. Knowledge convergence is based on Latent Semantic Analysis (LSA) and shows how similar one's language becomes to their team's language over time. Following the steps outlined in Dong [35], LSA is used to identify the topics a team discussed, and each team member's cumulative communication is compared to these topics to identify when new topics are added to the conversation and by whom. We chose not to label teammates on this graph to preserve privacy.

Semantic Coherence: Semantic coherence also comes from Dong [35], and is executed similarly to knowledge convergence. However, semantic coherence is a team-wide, non-cumulative measure, calculated as the norm of the sum of each team member's LSA vector per day.

Embedding Space Similarity: The last commonly used measure of shared understanding in text uses embedding spaces [24, 78, 94], which transform words and sentences into vectors that retain meaning. We use the method defined by Lix et al. [78], where we calculate the cosine similarity between embedding vectors for each pair of team members per day and average this similarity across all pairs. For all graphs, when large time periods are chosen, we apply a 3-day moving average to smooth the visualization before displaying it.

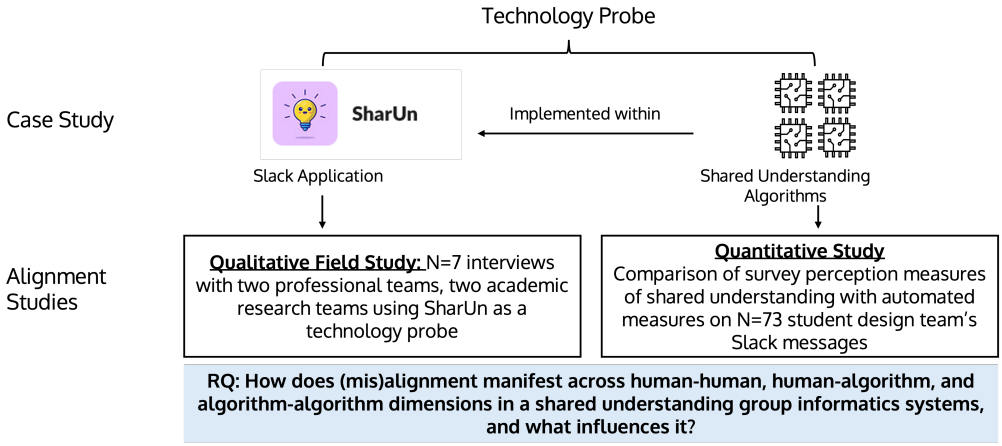

Figure 2: An overview of the study design. SharUn is implemented as a technology probe, focused on measuring shared understanding within teams. In the qualitative study, seven members from four teams used SharUn with their own Slack workspace during the interview. In the quantitative study, the algorithms are applied post-hoc to engineering design teams' Slack messages, and compared to their survey responses of perceptions of shared understanding.

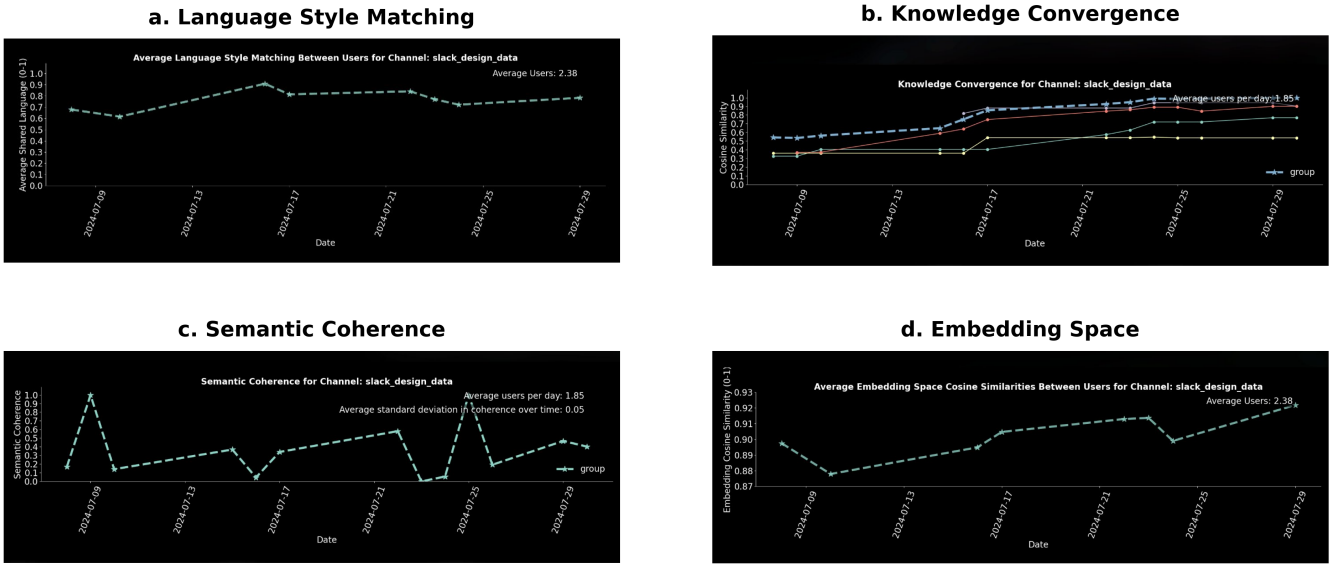

Figure 3: Example of the various algorithmic measure graphs generated by SharUn. Calculated for one channel from the authors' Slack workspace. (a) Language Style Matching (LSM) measures similarity in language style based on the frequency of function word use [15] in user pairs (higher values = more similar styles). (b) Knowledge Convergence shows individual alignment (unlabeled lines) with team-wide topics (dashed line) via Latent Semantic Analysis, which compares the number of topics each member has discussed on each day with the team's total during the time period. (c) Semantic Coherence, also measured via Latent Semantic Analysis, captures daily team-wide topic similarity based on word choice, meaning that similar topics are identified when similar words are used. (d) Embedding Space Similarity measures thematic alignment between user pairs via cosine similarity of embedding vectors. This measure captures teams discussing similar topics using different terms.

3.2 In-Situ Qualitative Study

We interviewed seven participants from two academic and two professional teams (Table 1), reflecting long-term team contexts. We used SharUn as a functional technology probe to calculate metrics from the team's conversations, allowing participants to reflect on and compare their interpretations of specific conversations with the algorithmic results.

3.2.1 SharUn Technology Probe. SharUn was designed as a technology probe to investigate the alignment between team members and the various algorithmic measures of shared understanding. Technology probes are functional systems deployed in real-world settings to understand user needs and inspire new technological possibilities [14, 60]. Functionally, SharUn measured each team's shared understanding and visualized this information in a dashboard for participants to reflect on alignment (see Figure 4). Within the probe, we applied four traditional shared understanding algorithms to understand the extent to which each algorithm represented the participants' perceptions of their shared understanding. The algorithms captured various approaches to measuring shared understanding (Section 3.1) and enabled the analysis of message content in real-time, which is required to reflect on alignment.

To support understanding of the algorithmic approaches and facilitate reflection, we designed the system to visualize shared understanding patterns at the team level over a chosen period of time. We provide static descriptions of each shared understanding measure (see Figure 4). As a technology probe, SharUn prioritized enabling reflection on perceptions of alignment between their own understanding and the algorithmic results, which was critical to our investigation. The probe was implemented as a Slack application (see Figure 4). Chat-style messages contain information about the team's collaboration patterns [18, 46], work tasks [45], and support the process of building shared understanding [12, 88]. For privacy, the app could only view content from *public* Slack channels to which it was explicitly added. Once added, the system automatically sends opt-in consent forms via direct messages to every channel member. Messages are analyzed when the user visits the app to visualize their collaboration. On the app's homepage, the user can select a channel and a time frame to visualize their shared understanding. We used the four algorithms previously described to estimate shared understanding. Once the algorithms had been applied to the filtered input data, four shared understanding estimates over time would be displayed in graphs shown to the user. The data flow is shown in Appendix B.

3.2.2 Participants. We recruited n=7 participants from four teams for the in-situ study (Table 1). We recruited participants through a mix of sampling the researchers' networks, as well as posting on social media (X and LinkedIn). We only required one participant per team, but we encouraged participants to ask others on their team if they were interested in participating to capture alignment from multiple viewpoints. The teams had experience working together for at least 4 months prior to the study, with the most senior team having worked together for 3 years. Many studies suggest that shared understanding can develop over the course of a few months, and researchers study shared mental model development over the course of a semester [77, 91], or, even over the course of less than an hour [100]. Participation was not compensated, and all team members who were part of the analyzed Slack channel, as well as the participant interviewed, provided informed consent (University of Toronto Research Ethics Board #: 41433)

3.2.3 Study Design. A few days before data collection began, each team downloaded the application and added it to at least one public channel. Then, we conducted one-hour interviews with 1-2 members from each team. We asked participants how they would identify shared understanding (or its absence) when collaborating with their team. Then, participants used the probe to generate graphs representing their teams' shared understanding for a specific time period (we recommended a two-week period with extensive discussion) in a Slack channel. We then asked them to review the dashboard before speaking aloud their thoughts and interpretations. Participants were encouraged to read the static interpretations of each measure presented on the app (Figure 4), but no other interpretation guidance was provided. We asked participants to identify 2-3 points of interest on one or more graphs and navigate to that time period in the specified channel. We asked participants how their interpretation aligned with or did not align with what the graph would suggest (as periods of high or low shared understanding), their own thresholds for high or low shared understanding, and their confidence in the measurements.

3.2.4 Analysis. We qualitatively analyzed the resulting interview data using a hybrid thematic analysis approach [43]. We guided our analysis based on the three types of alignment (Human-Human, Human-Algorithm, Algorithm-Algorithm). After the first author collected transcript segments for each alignment type, we conducted a bottom-up thematic analysis [16] to identify an initial set of codes for each alignment type. In collaboration with the last author, these codes were grouped into axial codes that describe each type of alignment at a high level. Disagreements were resolved through consensus discussion, often returning to the original quotes for verification. To establish trustworthiness, we used reflexivity memos and negative case checking by returning to original quotes to verify or challenge emerging interpretations.

Within Human-Human alignment, we identified 13 codes describing shared understanding, which were then collaboratively grouped into 5 higher-level categories (e.g., linguistic markers of misunderstandings, diverse language). By tabulating how many participants used each code, we identified which descriptions were widely used. In Human-Algorithm alignment, bottom-up analysis resulted in 21 individual codes. These codes were grouped into 10 higher-level reasons for perceptions of (mis)alignment, including recognizing key events, recognizing general trends, and a lack of expected variation. Lastly, for Algorithm-Algorithm alignment, we initially identified eight codes describing how participants discussed the different graphs shown. From these codes, four high-level categories were developed (e.g., higher is better, complementarity of measures) that describe perceptions of A-A alignment.

Next, three researchers collaboratively conducted a cross-category thematic synthesis across all three alignment types. We placed all higher-level categories on sticky notes, and physically grouped these categories. We iteratively revisited the quotes within each code group to generate a higher-level alignment finding. We would then check whether the themes cross-cut different alignment types, adjust as needed, and revisit the quotes. Throughout this process, we used reflexivity memos to record how our themes and code groupings evolved. We started with four themes (multiple complementary measures, H-H alignment as a barrier to H-A alignment, importance of visual presentation, and the converging influence of key events). Through consensus, we consolidated these themes into two: the complexity of team dynamics and the influence of key events. We present key findings (subthemes) within these themes.

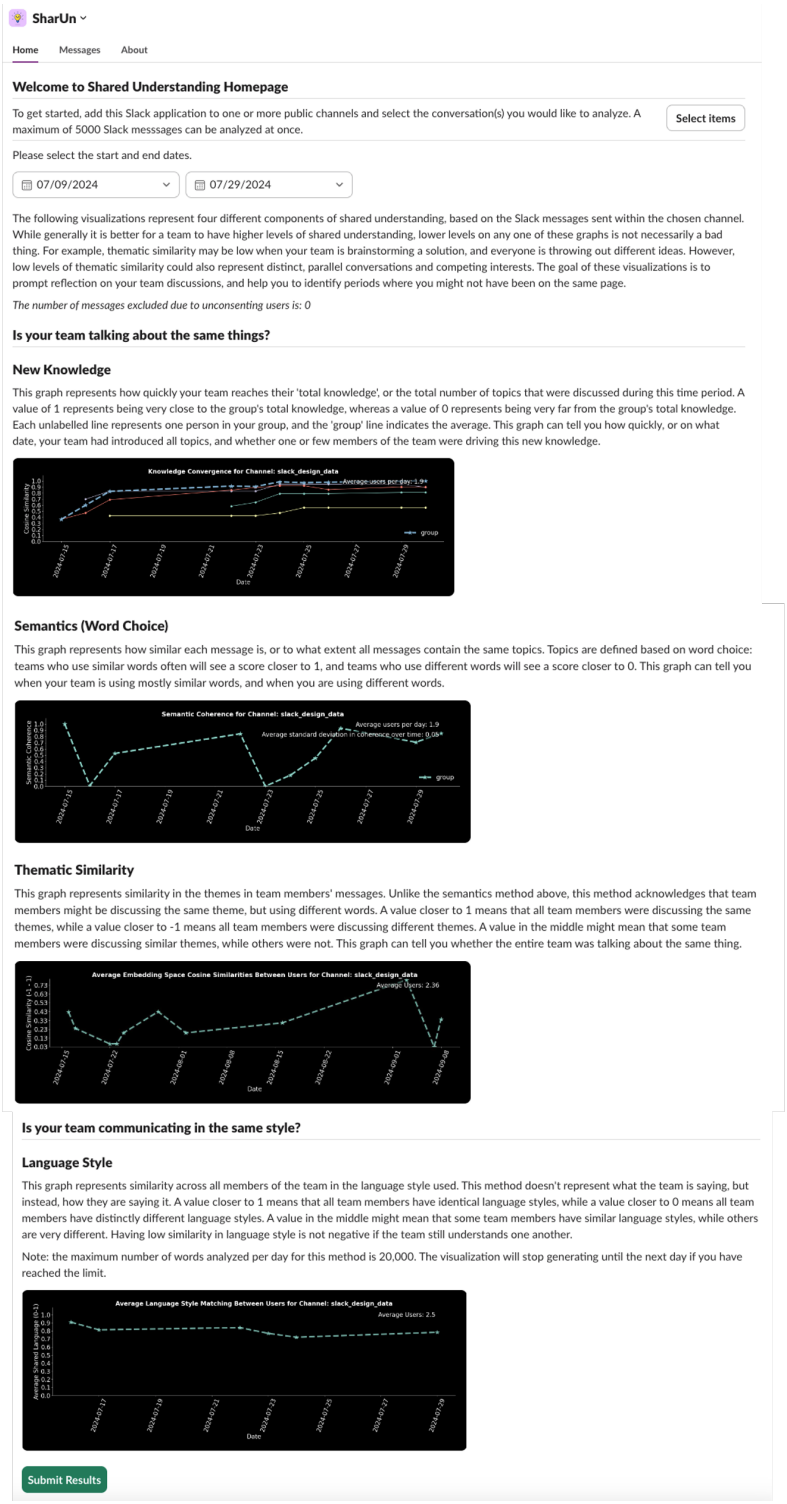

Figure 4: Complete interface of the SharUn homepage. Includes the interpretive explanation for each graph.

3.3 Quantitative Study

3.3.1 Participants. For the quantitative study, we studied a fourth-year engineering design capstone course. Engineering design is a

Table 1: List of participants interviewed in the pilot evaluation. A - Academic team, I - Industry team

Participant	Workspace	Organization Type	Project
P1	A1	Research Lab	HCI Research Project 1 channel
P2	A1	Research Lab	HCI Research Project 2 channel
P3	I1	Software Development	Team-wide general channel
P4	I1	Software Development	Organization-wide product channel
P5	I2	Software Development	Team-wide general channel
P6	I2	Software Development	Team-wide general channel
P7	A2	Research Lab	Education Research Project 1 channel

rich setting to study computer-mediated communication [48] and shared understanding, as it is a team-based discipline in which members often have non-overlapping expertise [93]. The six teams studied are comprised of 14–16 students who work alongside 10–15 staff members for 3.5 months. Despite the educational setting, this environment presents more realistic teamwork challenges than short-term teams, due to the interdependence and the need to manage their own schedules and budgets. The teams are hybrid with regular in-person meeting times, and they often communicate via Slack when the team members are not together.

N=27 participants participated in the quantitative study (e.g., those who shared their Slack data and completed the surveys). Of these, 12 identified as men, 12 as women, and 3 preferred not to say; Asian (n=6), Hispanic or Latino (n=5), and White (n=15) were the primarily identified races. All participants had worked closely in a team for at least one month by the first survey, and three months by the final. With consent, we collect all public channel Slack communication from these teams. Past HCI work has also used realistic education settings to evaluate algorithms [23].

3.3.2 Study Design. To compare survey perceptions of shared understanding, we surveyed these teams three times throughout the project (beginning, midpoint, and end). We used two constructs from the high-quality relationships measure from Carmeli and Gittell [20] (Appendix C, shared goals ($\alpha = 0.66$) and shared knowledge ($\alpha = 0.66$), measured using three and four questions, respectively. Each question was scored on a 5-point Likert scale.

To evaluate the algorithms, since each measures shared understanding per channel, we average the shared goals and shared knowledge responses for all active team members (i.e., those who sent at least one message) in that channel over the three days preceding the survey. We compare this to the 3-day-averaged values of shared understanding for these teams, based on each algorithmic measure. We compare the results from each algorithmic measure using Spearman's correlation. The knowledge convergence algorithm was excluded from this analysis because it is the only cumulative measure and it would not be expected to correlate with the non-cumulative measures (see the resulting correlation matrix in Figure 5). For the LSA-based algorithms, since the results depend on the chosen time period, we calculated the algorithms separately across three time periods aligned with the survey distribution. This quantitative investigation was conducted post-hoc, and entirely separate from the technology probe study. Participants were not shown any visualizations, or the results of the applied algorithms.

4 RESULTS

Across both studies, our analysis revealed two major themes. First, we found that achieving alignment in group informatics systems is fundamentally complicated by the multidimensional nature of teamwork constructs: algorithms capture different facets of shared understanding, human perceptions do not map consistently onto any of these facets, and visual framing of algorithmic output further shapes interpretation of alignment. Second, we found that the alignment that does exist is concentrated at extremes and key events, with day-to-day variation largely going unrecognized.

4.1 Complexity of Team Dynamics Requires a Holistic Measure

One of the two major themes we identified is that teamwork constructs are perceived as multidimensional and complex. Across the three types of alignment, we surface three specific findings. First, the algorithms capture different facets of shared understanding and are perceived as complementary. Second, human perceptions of shared understanding do not map consistently onto any single measure, creating a barrier to Human-Algorithm alignment, as team members themselves disagree on what shared understanding entails. Third, visual presentation of results systematically shapes how users align.

4.1.1 Algorithms Capture Different Facets of Shared Understanding. **Quantitative Findings:** The shared understanding algorithms we implemented showed substantial misalignment, suggesting there is no consensus approach to measuring shared understanding. With the student population, we quantified alignment using correlations. After applying each algorithm to the design teams' Slack messages, which yielded one value per team and day (6 teams, approximately 96 days), we calculated Spearman's correlation for each measure. As mentioned previously, the knowledge convergence algorithm was excluded from this analysis. The correlation matrix is shown in Figure 5. Although all correlations are significant due to the large sample size ($p < 0.001$), the only notable moderate-strong correlation is between semantic coherence and embedding space similarity metrics, with a negative relationship. This suggests that the proposed algorithmic measures of shared understanding do not align with each other.

Qualitative Findings: Our qualitative study with professional and academic teams suggests that participants perceived the algorithms as complementary — they capture distinct facets of shared understanding. Most participants recognized that the algorithms did not always align with one another. As participants naturally

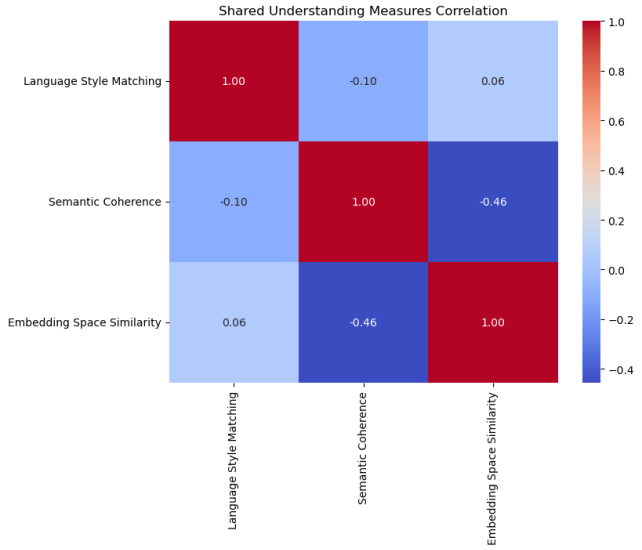

Figure 5: Spearman's Correlation matrix comparing three algorithmic measures of shared understanding (language style matching, semantic coherence, and embedding space similarity)

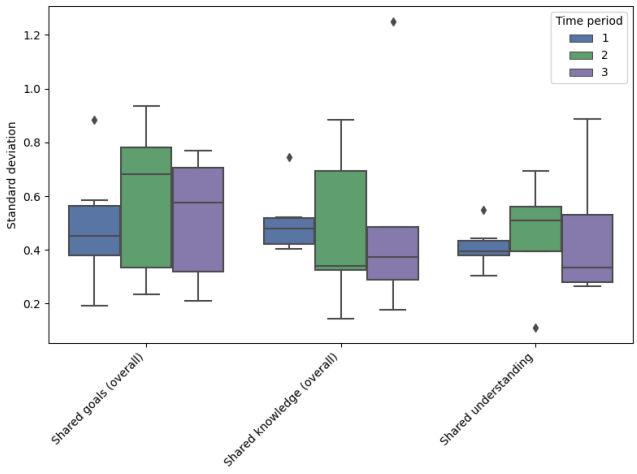

Figure 6: Range of per-team standard deviations for each component of shared understanding, per time period.

compared the four graphs, they expressed interest in seeing each measure overlaid for easy comparison: *"it would be kind of cool to overlay the two...So if we're talking about the same themes, but using different words for it...I think the discrepancies are perhaps more interesting.."* [P2]. The complementary nature promoted deeper reflection:

> "I read each of them differently, and then I took a quick glance through [them] together..That definitely made me notice much more larger trends that I wasn't looking at when I was just going through each of them individually, but I think both comparisons...felt helpful [in] reflecting upon it..." [P7]

Participants expressed preferences, clarity, and confusion about different metrics—they showed no alignment on which of the four metrics might be most promising. For instance, one participant expressed confusion about the value of the language style metric, *"language style...is that good? Or is it bad? Is it good because we're [a] diverse team? Or is it bad because we're misaligned?"* [P1]; while another noted that this was the most interesting metric *"I find the last one to be pretty helpful, or it's more of an interesting data point...I don't know enough about this to know whether style matching has an impact on shared understanding. But it is interesting"* [P3]. Sometimes, this misalignment occurred within the same participant, with P2 stating that the knowledge convergence graph made sense as it aligned with different project phases: *"if it's more process related stuff, then you would probably want the lines to be closer together. But if it's more ideological tasks where diversity and then different perspectives are a better thing, you'd probably want to see more of that distance..."*, yet later said that the cumulative nature was challenging to interpret, *"I don't know if I fully understand it yet, but I understand it's time and...how many topics were being introduced. So it's a little bit harder to, I think, contextualize in context of the conversation"*. Participants frequently compared the semantic coherence plot with the embedding space model, though disagreed on their interpretation: P4 notes *"I don't understand, the definition of a theme, but like I do understand the definition of a word"*, yet P3 says ' *'the average embedding space cosine similarities...I think this one is probably the most useful just to see people kind of stay on the same topic."*

The details from the qualitative study suggest that humans recognize shared understanding as a multifaceted construct that may be captured by multiple, complementary approaches.

4.1.2 H-H Misalignment as a Barrier to H-A Alignment. First, we quantify variation in teammate perceptions of shared understanding from the survey measures in our student population. Then, we provide examples of these different perceptions from our qualitative study.

Quantitative Findings: From the survey perceptions, we calculated the standard deviation for each question, within each team and time period. These values ranged from 0 to 2.12, with an average of 0.6 (on a 1-5 scale, see Appendix D). Particularly when positive response bias is considered (participants rate their team highly—the overall average shared understanding score was 3.9/5), this indicates moderate disagreement among team members regarding their shared understanding (see Figure 6).

Our quantitative analysis further revealed a gap between what teams perceive as shared understanding (measured via surveys) and what current algorithms measure over time. As surveys cannot be collected as frequently as digital trace data, we are comparing data at different levels of granularity (daily vs. monthly). There is more variability in the algorithmic measures than in the survey measures—and the automated measures vary cyclically over time (see Figure 7, which displays algorithmic measures and survey responses for one Slack channel). This was also shown when measuring discursive diversity in teams over time [78]. Due to this variation and the recall bias in surveys [104], we compare

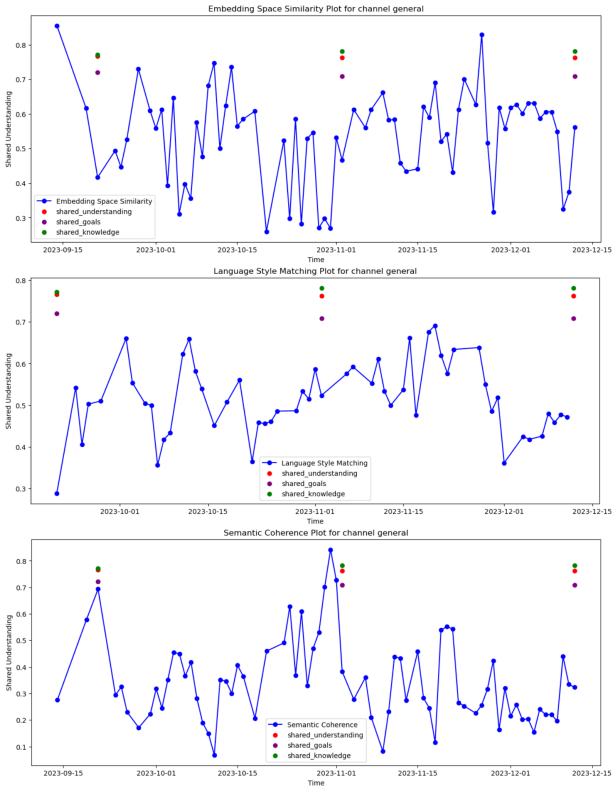

Figure 7: Example plots from the #general channel from one student team showing the variance in each algorithm over time (lines), compared to variance in survey responses at three time periods (points). Note that the surveys (measured on a scale of 1-5) are scaled between 0-1 for this plot.

the average of the algorithmic metrics for each channel in which the participant was active *three days prior to the survey* with the reported survey values.

We calculated this value for shared goals and shared knowledge, separately, as well as their average, shared understanding, shown in Table 2. These correlations are calculated based on n=73 returned surveys (out of 83, where 10 surveys were removed) for not responding to all questions (n=4) or for having no corresponding algorithmic results on the days prior to the survey (n=6).

Table 2 shows that there doesn't seem to be a correlation between the survey measures of shared understanding and the algorithmic measures, with only four significant *negative* correlations. In fact, the average of the three shared understanding metrics is significantly *negatively* correlated with shared understanding and shared goals. While there is a lot of variation in the algorithmic measures, this signals a disconnect between how team members may interpret shared understanding and the similarity they display in their textual communication. Thus, the human-human (H-H) misalignment seen in our data suggests a fundamental challenge: because people disagree on what constitutes shared understanding, it may be very difficult to create human-algorithm (H-A) alignment at all.

These standard deviations and negative correlations provide a quantitative signal of Human-Human disagreement, but cannot explain why disagreement occurs—the in-situ study fills this gap.

Qualitative Findings. This is further illustrated in the qualitative study as participants shared varying descriptions of how high or low shared understanding would manifest in text. One participant described repetition as a signal of low shared understanding: *"where is the data set? like a few times...the data set is at the top of the channel"*[P1]. Others said that explicit negations or disagreements indicated low shared understanding (one participant): *"no this isn't part of our roadmap"* [P4]. Still, other participants said multiple questions being asked was a marker of misalignment (two participants): *"there's kind of a misalignment [and] I'll be like, Hey, are we meeting?"*[P2]. We did see some alignment within the members of team I1, where P4 perceived low shared understanding as negations, and P3 perceived high shared understanding as agreement (as did two other participants): *"everyone's kind of agreeing. No one's saying, I don't know..'* [P3]. Interestingly, none of these signals were explicitly captured by the algorithmic measures we applied, which focus on consistency in terminology and context. And so, team members not agreeing on what shared understanding looks like in the first place presents a core challenge for Human-Algorithm alignment, making it difficult to design team-level measures that satisfy a definition that does not exist at the team level. The consistency between these descriptions of Human-Human alignment and the patterns observed in the student teams suggests that divergent interpretations of shared understanding may exist across contexts.

4.1.3 Visual presentation shapes interpretation independently of algorithmic content. **Qualitative Findings.** Beyond the content of algorithmic measures, we found that the visual presentation of results shaped how qualitative participants interpreted shared understanding, often overriding the guidance provided by the system and researchers.

Despite various interpretations of the construct, we consistently found that the visual presentation of the data also influenced the interpretation of shared understanding metrics. Even with disclaimers on the SharUn homepage that high graph values were not always positive, and low values were not always negative, participants consistently aligned on this interpretation: *"So top of the graph is good...I'm looking at this graph and I'm saying...I think we're not bad, because we're mostly at the top of the graph"* [P1]; *"I would say...each individual is very close to one...So it means...if it's closer to one...the team has a good knowledge of that topic"* [P6]. For the knowledge convergence graph, which displays each team member as a separate line on the plot, participants interpreted converging lines as a positive collaboration signal: *"for most users..we're pretty close already to the total knowledge... "* [P5]. Some participants took a more nuanced approach in their interpretation, noting that diversity in discussion topics or language style is not necessarily a bad thing, or a signal of low shared understanding: *"I don't place a lot of value on [the] kind of the word choice people use...There's a lot of different words, just in everyday conversation"* [P3].

We observed that the anticipated misalignment between various stakeholders and how they would interpret the data lead to privacy concerns. Participants were reluctant to share these dashboards with people who lack context about the data: *"it is flagging that*

Table 2: Spearman's correlation of each survey construct with each automated measure of shared understanding

Survey Measure	Language Style Matching	Semantic Coherence	Embedding Space Similarity	Averaged Metric
Shared Understanding	$r = 0.0, p = 0.97$	$r = -0.20, p = 0.08$	$r = -0.25, p < 0.05$	$r = -0.24, p < 0.05$
Shared Goals	$r = 0.41, p = 0.73$	$r = -0.24, p < 0.05$	$r = -0.19, p = 0.10$	$r = -0.24, p < 0.05$
Shared Knowledge	$r = 0.01, p = 0.9$	$r = -0.11, p = 0.33$	$r = -0.22, p = 0.06$	$r = -0.17, p = 0.15$

were misaligned. So that's a bad thing... " [P1] They also did not want to share their data with senior leadership, for fear of their shared understanding data being misinterpreted: *"...for people who work in the team this can be very helpful, but I'm not sure what happens if it...goes to senior leadership"* [P6]. P6 would not want senior leadership viewing these statistics because they worry that they are not aligned with what the values mean, and whether they are positive or negative. These concerns were rooted in the anticipated human-human (H-H) misalignment—participants worried that without a shared understanding of the metrics, the data could be misinterpreted.

4.2 Alignment is Influenced by Key Events or Extremes.

Qualitative Findings. Our second major theme is that alignment is influenced by key events or extremes. This theme is supported only by our qualitative study, where participants reflected in real-time on the recent extreme events their team had experienced. We saw increases in both human-human (H-H) alignment and human-algorithm (H-A) alignment during extreme or unexpected events. There were two instances in which participants were able to guess what happened on a specific day by viewing the graphs. For example, P6 noticed a spike in many graphs, and suggested that this was likely an incident or urgent issue with the team's software product: *"So I'm assuming there was an incident on this day that everybody was talking about the same topic and trying to figure out what was going on."* Upon investigating this day, they confirmed that there was a technical difficulty affecting all engineers, requiring collective action. This also occurred when P7, a research student, saw the graph for their team—they were able to immediately identify the day of a big presentation due to spikes in their conversational similarity: *"I can...definitely see why this is...kind of higher similarity...before that super big presentation, and then we hit the day of the presentation...We stopped talking about the presentation, and we all started doing our own things again.."*

Participants often described instances of perceived alignment as out of the ordinary. For example, P5 describes how the team all used a specific short-hand to refer to a meeting on a day where the semantic coherence score was high: *"it's interesting [the meeting] has a longer name. But everyone...had...the same shorthand."* Other participants described how brainstorming solutions resulted in low embedding space similarity, as would be expected, *"...the product team was doing a little brainstorming on the best way to address it..."* [P4]. Some participants interpreted the graphs in the context of the channel being analyzed. As P5 analyzed a team-wide channel, they described how they expected lower similarity than they would find in a project-specific channel: *"It makes sense to me that [the line] is in the middle...amongst the team members, not everyone is required to discuss the same thing...Across seven developers... they're working on different tickets"* [P5]. These findings suggest that H-A alignment may be higher during extreme, memorable, or uncommon events (when shared understanding is very high or very low).

When explaining the perceived alignment during these extreme events, participants often referenced similarity in word choice, reflecting the semantic coherence measure: *"...seems like there's a lot of similar words in those 2 messages"* [P2]; *"like all the research teams that were on campus called the Lunch and learn, and at least half of these messages, mentioned the lunch and learn"* [P7]. Participants also identified alignment with the language style matching measure, based on their general understanding of differences within their team: *"the fact that our roles are very similar. And...educationally and personality-wise people are very similar in the team. So...they can use similar language style"* [P6]. Participants discussed thematic similarity in passing, but did not often use this graph to describe their alignment with extreme events. The knowledge convergence metric was never used in explaining alignment at extreme events.

In contrast, participants did not always perceive the subtle nuances of shared understanding captured by the algorithms. The graphs showed regular ups and downs in shared understanding metrics on a day-to-day basis, but participants did not always recognize these. When participants described instances where the algorithms did not align with their perceptions of the conversation, they described misalignment with their general perception of the team (*"It's kind of weird how they are almost always the same...there's no variance at all"* [P3]); misalignment with linguistic markers of shared understanding (*"It was a moment of discussion, but not a moment of disagreement"* [P1]); or misalignment with thematic evolution throughout the conversation (*"the conversation evolved...So [person is] asking for A, we're seeing if we can do B....And then he says, No, that's not helpful...And so we start talking about C."*). Notably, the algorithmic measurements captured instances of extremely high or extremely low levels of shared understanding, aligning with participants' interpretations, but the subtle daily fluctuations went largely unnoticed.

5 DISCUSSION

In this work, we provided an initial, exploratory characterization of how misalignment manifests across three dimensions (Human-Human, Human-Algorithm, Algorithm-Algorithm) in group informatics systems. Across a qualitative in-situ study and a larger-scale quantitative evaluation, we found little evidence of alignment, and provide initial evidence to explain why this alignment is absent. Two key themes emerged regarding what affects alignment perceptions: teamwork constructs are perceived as multifaceted and complex, and alignment increases during extreme events. Within the theme of construct complexity, three key findings stand out: low A-A alignment might indicate complementary facets of shared understanding; H-H misalignment acts as a prior barrier to H-A

alignment; and visual framing of algorithmic outputs can systematically shape alignment. We contribute an exploratory empirical account of how misalignment manifests across all three dimensions, informing the emerging discussion of group informatics systems, what they should be designed to measure [79], and what ethical challenges and considerations will arise [73]. Here, we discuss open questions and future work related to constructing holistic measures of team dynamics, the value in capturing day-to-day variation, and whether alignment is feasible and truly necessary for useful systems.

5.1 Co-Constructing Holistic Measures of Team Dynamics

Across the two populations, our qualitative and quantitative findings provided initial evidence of misalignment between team members and algorithms. The algorithms had little correlation with one another, except for the semantic coherence and the embedding space similarity models, which were negatively correlated. This may be due to the tendancy for new teams to use different words for the same concept [50]; for example, someone may say "the document", while someone else refers to the "word file". In this scenario, the thematic similarity would be high, while the semantic similarity would be low as there are no shared words. On the other hand, for teams that have been working on a project for a while, they may develop specialized short-hand that may be consistently used and thus captured by semantic similarity models, but may not be included in the dictionary and thus left out of embedding space models. Whereas semantic similarity captures surface form and thematic similarity captures meaning, it is not surprising that these may not always move in the same direction, as they capture shared understanding at different levels of abstraction. The lack of correlation between language style matching and semantic coherence may also be explained by past work that showed that topic modelling based measures increased with highly involved interactions, whereas language style matching increases with emotional conversations [7]. Our qualitative study found no preference for one algorithm over another, though participants identified alignment at extreme events primarily though the language style and semantic similarity graphs. While we interpret this misalignment as evidence of construct multidimensionality, future work can empirically test the impact on performance depending on which algorithm's results are presented to users.

We identify specific findings that complicate alignment and carry distinct consequences for group informatics systems: a sequencing problem, a visual interpretation influence, and varying alignment across events. The low A-A alignment may not indicate algorithmic failure, as qualitative participants perceived divergent algorithms to capture complementary facets of shared understanding. Instead, this may be further evidence that team dynamics are multidimensional [1, 71] and challenging to capture algorithmically [54].

A simple answer to this challenge might be to collect larger samples of definitions for each construct (e.g., through crowdsourcing) to design systems that integrate theoretical models and workers' understandings. Yet Balaam et al. [9] argues that we should not try to formalize something that cannot be formalized—perhaps, for example, a strict definition of successful teamwork. There is a lot to gain by translating academic theories to automated measures, such as the ability for teams to receive consistent feedback. But this process is not always straightforward, as the theoretical definitions are not always aligned with team members' perceptions of the constructs in practice, raising a tension between theoretical best performance and user acceptance. For instance, even though the algorithms used here did not align with user perceptions, Language Style Matching still predicts team cohesiveness [52], cycles of high and low semantic coherence are associated with higher performance [101], and greater diversity in embedding space similarity relates to team performance [78]. Perhaps algorithm selection for group informatics should be driven by the construct facets relevant for a given team context, instead of a single measure.

Our findings also point to a few directions that might lead to more effective measurements of shared understanding. Our qualitative participants identified signals of shared understanding, such as explicit agreements or disagreements, questions, indications of confusion, and repetition, which were not explicitly captured in existing algorithms. Future work might explore dialogue act modelling [102] to identify the intent of messages, such as acknowledgments, repairs, or question patterns.

Systems may also enable teams to co-construct holistic measures of complex teamwork constructs. Aligning with collective sensemaking theories [65, 92], group informatics systems could support the ongoing social process of negotiating metrics. The inherent variation in members' interpretations of complex teamwork constructs, coupled with their multidimensional nature, means these systems may need to present multiple metrics while balancing information overload. We noticed that the in-situ participants expected that shared understanding could not be captured in a single measure, and defaulted to comparing each graph, using discrepancies to prompt deeper reflection. These holistic measures may also incorporate self-reports and reflections. Future studies of group informatics systems should empirically evaluate this balance between capturing holistic measures, increasing H-A alignment, and information overload.

We also found that visual presentation of results systematically shaped how participants interpreted shared understanding, overriding guidance from the system. Despite the variance in how qualitative participants described shared understanding, one interpretation seemed consistent: that high values on the graphs were "good", and low levels were "bad." Participants seemed most worried that people outside of their immediate team might be misaligned on what was "good" or "bad" and would judge these statistics without knowing the context of their collaboration. As such, the visual design of a graph may have influenced consistent interpretation—future design of group informatics should investigate how visual design can constrain interpretation, and in which situations this is ideal.

5.2 Alignment at the Extremes

Our qualitative, in-situ study revealed that Human-Algorithm and Human-Human alignment existed primarily during "extreme" events in shared understanding—participants noted that during unexpected events, such as a software outage, or uncommon events, such as team-wide brainstorming, the algorithms displayed the expected levels of shared understanding (high and low, respectively).

This aligns with the literature, which shows that sensemaking is often triggered by disruptions or uncommon events [80, 110]. Yet the daily fluctuations between these extremes confused participants who perceived that their team generally had relatively stable levels of shared understanding outside of these events. As such, if teams do not perceive routine variations as problematic, they have no reason to collectively interpret them.

We raise the question of whether alignment in day-to-day variation (or agreement on whether this variation exists), matters. If group informatics are designed to actively provide feedback to teams, it is likely that this feedback would be most useful during extremes—particularly extreme lows in shared understanding—where we see higher levels of H-H and H-A alignment. Participants also feared this routine variation being misinterpreted by outsiders. Thus, perhaps group informatics systems should focus on identifying the extremes. This could be done through event-triggered or anomaly-based approaches, selectively surfacing moments where the measures deviate significantly from a baseline.

On the other hand, extreme teamwork failures may be easier for teams to identify themselves. Daily variation in shared understanding may also be meaningful, as persistent low shared understanding levels could be a potential predictor for future intra-group conflict that team members may not yet recognize. However, it is important to take a nuanced look at shared understanding— consistently high shared understanding can also discourage constructive divergence between team members, potentially stifling creative ideas and innovation [29, 99]. Some variation in shared understanding is likely valuable, but the optimal range remains an open question that warrants further investigation as we continue to develop SharUn and collect more data on team shared understanding and outcomes. Bringing teams' attention to subtle inefficiencies in their collaboration could make them more effective collaborators in the long term, yet granular monitoring risks alert fatigue and, more concerning [17], enables granular surveillance of daily behaviours that teams themselves don't consider problematic.

5.3 Dealing with Misalignment

Our findings suggest that misalignment in group informatics systems operates at multiple levels. Algorithm-Algorithm misalignment may not need to be resolved if users interpret different measures as complementary. However, the lack of Human-Human alignment is more fundamental, suggesting that group informatics systems will be evaluated against different standards by different team members. The visual interpretation finding suggests that even well-designed algorithmic measures may be systematically misread in practice. We also note that our sample was modest, so H-H alignment within teams could differ with added team members. This also means that we cannot necessarily determine whether the algorithms align with the team's negotiated understanding of their shared understanding. However, our work demonstrates that people within and between teams have differing interpretations of shared understanding, meaning that teams will likely struggle to converge on one perception of it, and this will likely not be captured by algorithms.

Fundamentally, our results raise the question of whether alignment is achievable, or even necessary. We saw evidence of misalignment across all three types of alignment evaluated. Our quantitative study revealed little correlation between algorithmic measures; our qualitative participants noted that they found value in divergent metrics, using discrepancies to deepen their understanding of the construct. Forcing algorithmic alignment—by choosing a single measure—would eliminate this richness. Addressing H-H and H-A misalignment, group informatics systems could support individualized representations of the data. For example, Balaam et al. [9] introduced a system that enabled schoolchildren to communicate their emotions to their teacher. They allowed students to choose how to represent their emotions (via colours), and they allowed the teacher to set individualized rules for what constituted a positive or negative emotion, customizing their dashboard to align with these rules. Similarly, group informatics systems could allow team members to personalize their dashboards according to their interpretation of successful teamwork. While this might work for systems that capture individualized measures of collaboration, it is counterproductive for systems that capture *group-level* measures. To improve upon these measures, teams would need to take coordinated action, and collaboratively discuss the measures—which would be challenging if teammates saw different dashboards.

Previous work suggests that perfect alignment should not be the goal, and that instead we should aim for functional alignment in measures that teams find useful for reflection [68], even if they do not perfectly match every member's intuition. This raises the question of who the systems should be useful for. It is well known that workplace data collection and reporting are common forms of workplace surveillance [53, 55], which makes 'useful for whom' a political question. Even measures designed to support team reflection may be misused to exert control over employees. We argue that the question of alignment is important for distinguishing teamwork feedback systems from surveillance machines. Ultimately, whose interests are served by metrics depends on who controls their definition and interpretation [86]. These systems have the potential to change who is observing work [54], who has authority [86], and who can set normative expectations [54, 86].

5.4 Limitations and Future Work

The insights from this study should be interpreted with some limitations in mind. First, we focus our discussion of group informatics systems as tools for knowledge workers, yet past work has shown that workplace measurement is applied differently to different types of work [55]. While our sample size for the in-situ, qualitative evaluation was small, we were able to identify a breadth of perceptions of shared understanding, and differences across members of the same team. The in-situ user study focused on a single team channel to allow participants to discuss alignment across different points in a single conversation, but prevented us from assessing whether alignment perceptions differed by conversation topic.

Our technology design probe was designed to enable participants to reflect on their perceptions of shared understanding, and whether the algorithms aligned with these perceptions. Thus, the goal was not to measure the effectiveness of such a tool in the long term, and we leave the question of the effectiveness of group

informatics systems for future work. Particularly, we note that our qualitative study prompted participants to reflect on and critically analyze their shared understanding, which is a process that may not happen naturally, and likely will affect how beneficial the group informatics system is. Future work should examine which organizational conditions, such as how reflection happens and the team culture, are important to translate group informatics feedback into team action. While privacy and surveillance are incredibly important considerations in designing informatics systems, we have only scratched the surface of this discussion. While we note that some privacy concerns seem to stem from the assumption that there would be a misalignment in interpretation between workers and leaders, there are likely many more factors that influence perceptions of privacy.

The two population study characterizes alignment in text-based collaborative work across different populations, providing complementary perspectives on alignment. However, each population has limitations worth acknowledging. Each population, on their own, is modest in size—particularly limiting any statistical claims in the quantitative work. All qualitative participants are from academic or software contexts, meaning they may not reflect all knowledge workers, and may be more comfortable with algorithmic tools. The student teams reflect a more structured work process than the organic processes in professional teams.

In addition to the future research directions already mentioned, we share a larger group informatics research agenda. First, we plan to continue investigating other common challenges in informatics systems, such as privacy expectations in group-level measures and how to provide actionable, individual feedback when the group measure is more than the sum of its parts. We also foresee a future in which these systems can double as data-collection tools for researchers, enabling them to store anonymized, aggregated values and use them to study relationships between online actions, collaborative constructs, and team outcomes.

6 CONCLUSION

Social scientists have long known that successful team outcomes are related to a number of complex group-level constructs. While important, their nuanced nature and lack of measurement options make it challenging for teams to understand how they are doing and identify ways to improve. We argue that informatics systems can be designed to measure these group-level constructs, albeit with new and increasingly critical challenges. In this paper, we study one challenge in-depth: alignment. Using a functional technology probe that measures team-level shared understanding, we investigate the influencing factors of human-human alignment, human-algorithm alignment, and algorithm-algorithm alignment. We found little evidence of alignment across the board. We find that alignment is influenced by the need for a holistic measure of complex constructs, as demonstrated by the complementary nature of different measures, the human-human barrier to human-algorithm alignment, and the influence of visual presentation. We also find that alignment perceptions are influenced by extreme and unexpected events. We discuss the construction of holistic measures of successful teamwork, the importance of day-to-day variation in teamwork effectiveness, and the value of reaching alignment. We surface key design questions to consider as we continue to investigate the applicability of automated teamwork systems.

ACKNOWLEDGMENTS

Thank you to Kathy Cheng, Meagan Flus, Pranav Khanolkar, Marjan Naghshbandi, Chuma Asuzu, and Peiying Jian for feedback on early versions of this work. We are grateful for Sofiia Stelmakh's input and feedback on this work. Thank you to Receptiviti for providing in-kind support for this project through access to the LIWC API. This work was supported by a Microsoft Research AI and The New Future of Work grant. The authors wish to acknowledge the use of Perplexity and Claude AI for brainstorming possible phrasings. The paper remains an accurate representation of the authors' underlying work and novel intellectual contributions.

REFERENCES

[1] 2015. *Enhancing the Effectiveness of Team Science.* National Academies Press, Washington, D.C. https://doi.org/10.17226/19007

[2] 2023. In the Changing Role of the Office, It's All about Moments That Matter. https://www.microsoft.com/en-us/worklab/in-the-office-it-is-all-about-moments-that-matter

[3] Mohammed Almutairi, Charles Chiang, Yuxin Bai, and Diego Gomez-Zara. 2025. tAIfa: Enhancing Team Effectiveness and Cohesion with AI-Generated Automated Feedback. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*. ACM, Amsterdam Netherlands, 1–25. https://doi.org/10.1145/3729176.3729197

[4] Hayward P Andres. 2013. Team cognition using collaborative technology: a behavioral analysis. *Journal of Managerial Psychology* 28, 1 (2013), 38–54. https://doi.org/10.1108/02683941311298850

[5] Jorge Aranda, Ramzan Khuwaja, and Steve Easterbrook. 2007. Discovering the shared understanding dynamics of large software teams. In *Proceedings of the 2007 conference of the center for advanced studies on Collaborative research*. 244–247. https://doi.org/pdf/10.1145/1321211.1321237

[6] Mark S Avnet and Annalisa L Weigel. 2013. The structural approach to shared knowledge: An application to engineering design teams. *Human factors* 55, 3 (2013), 581–594. https://doi.org/10.1177/0018720812462388

[7] Meghan J Babcock, Vivian P Ta, and William Ickes. 2014. Latent semantic similarity and language style matching in initial dyadic interactions. *Journal of Language and Social Psychology* 33, 1 (2014), 78–88. https://doi.org/10.1177/0261927X13499331

[8] Sriram Karthik Badam, Zehua Zeng, Emily Wall, Alex Endert, and Niklas Elmqvist. 2017. Supporting Team-First Visual Analytics through Group Activity Representations. In *Proceedings of the 43rd Graphics Interface Conference (GI '17)*. Canadian Human-Computer Communications Society, Waterloo, CAN, 208–213.

[9] Madeline Balaam, Geraldine Fitzpatrick, Judith Good, and Rosemary Luckin. 2010. Exploring affective technologies for the classroom with the subtle stone. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1623–1632. https://doi.org/10.1145/1753326.1753568

[10] Peter Berggren and Björn JE Johansson. 2010. Developing an instrument for measuring shared understanding.. In *ISCRAM*.

[11] Tony Bergstrom and Karrie Karahalios. 2012. Distorting social feedback in visualizations of conversation. In *2012 45th Hawaii International Conference on System Sciences*. IEEE, 533–542. https://doi.org/10.1109/HICSS.2012.222

[12] Jeremy P Birnholtz, Thomas A Finholt, Daniel B Horn, and Sung Joo Bae. 2005. Grounding needs: achieving common ground via lightweight chat in large, distributed, ad-hoc groups. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 21–30. https://doi.org/10.1145/1054972.1054976

[13] Eva Alice Christiane Bittner and Jan Marco Leimeister. 2014. Creating Shared Understanding in Heterogeneous Work Groups: Why It Matters and How to Achieve It. *Journal of Management Information Systems* 31, 1 (July 2014), 111–144. https://doi.org/10.2753/MIS0742-1222310106

[14] Kirsten Boehner, Janet Vertesi, Phoebe Sengers, and Paul Dourish. 2007. How HCI interprets the probes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1077–1086. https://doi.org/10.1145/1240624.1240789

[15] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin* 10 (2022).

[16] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[17] Dan Calacci and Jake Stein. 2023. From access to understanding: Collective data governance for workers. *European Labour Law Journal* 14, 2 (2023), 253–282. https://doi.org/10.1177/20319525231167981

[18] Hancheng Cao, Vivian Yang, Victor Chen, Yu Jin Lee, Lydia Stone, N'godjigui Junior Diarrassouba, Mark E Whiting, and Michael S Bernstein. 2021. My team will go on: Differentiating high and low viability teams through team interaction. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–27. https://doi.org/10.1145/3432929

[19] Jean Carletta, Anne H Anderson, and Simon Garrod. 2002. Seeing eye to eye: an account of grounding and understanding in work groups. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society* 9, 1 (2002), 26–45. https://doi.org/10.11225/jcss.9.26

[20] Abraham Carmeli and Jody Hoffer Gittell. 2009. High-quality relationships, psychological safety, and learning from failures in work organizations. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 30, 6 (2009), 709–729. https://doi.org/10.1002/job.565

[21] Philip Cash, Elies A Dekoninck, and Saeema Ahmed-Kristensen. 2017. Supporting the development of shared understanding in distributed design teams. *Journal of engineering design* 28, 3 (2017), 147–170. https://doi.org/10.1080/09544828.2016.1274719

[22] Xinyue Chen, Lev Tankelevitch, Rishi Vanukuru, Ava Elizabeth Scott, Payod Panda, and Sean Rintel. 2025. Are We On Track? AI-Assisted Active and Passive Goal Reflection During Meetings. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–22. https://doi.org/10.1145/3706598.3714052

[23] Matt Chiu, Siska Lim, and Arlindo Silva. 2023. Visualizing design project team and individual progress using NLP: a comparison between latent semantic analysis and Word2Vector algorithms. *AI EDAM* 37 (2023), e18. https://doi.org/10.1017/S0890060423000094

[24] Matt Chiu, Arlindo Silva, and Siska Lim. 2022. Design Progress Dashboard: Visualising a quantitative divergent/convergent pattern of design team progress through Natural Language Processing. In *International Conference on-Design Computing and Cognition*. Springer, 67–84.

[25] HH Clark. 1991. Grounding in Communication. *Perspectives on Socially Shared Cognition/American Psychological Association* (1991). https://doi.org/10.1037/10096-006

[26] Courtney Cole. 2022. An Exploration of the Impact of Psychological Safety in Engineering Design Student Teams. (2022).

[27] Gregorio Convertino, Helena M. Mentis, Mary Beth Rosson, John M. Carroll, Aleksandra Slavkovic, and Craig H. Ganoe. 2008. Articulating common ground in cooperative work: content and process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1637–1646. https://doi.org/10.1145/1357054.1357310

[28] Gregorio Convertino, Helena M. Mentis, Alex Y. W. Ting, Mary Beth Rosson, and John M. Carroll. 2007. How does common ground increase?. In *Proceedings of the 2007 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '07)*. Association for Computing Machinery, New York, NY, USA, 225–228. https://doi.org/10.1145/1316624.1316657

[29] Nancy Cooke, Eduardo Salas, Janis Cannon-Bowers, and Renée Stout. 2000. Measuring Team Knowledge. *Human factors* 42 (02 2000), 151–73. https://doi.org/10.1518/001872000779656561

[30] Marianne Corvera Charaf, Christoph Rosenkranz, and Roland Holten. 2013. The emergence of shared understanding: applying functional pragmatics to study the requirements development process. *Information Systems Journal* 23, 2 (2013), 115–135. https://doi.org/10.1111/j.1365-2575.2012.00408.x

[31] Diane Coutu. 2009. Why Teams Don't Work. (2009). https://hbr.org/2009/05/why-teams-dont-work

[32] Vedant Das Swain and Koustuv Saha. 2024. Teacher, trainer, counsel, spy: How generative AI can bridge or widen the gaps in worker-centric digital phenotyping of Wellbeing. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*. 1–13. https://doi.org/10.1145/3663384.3663401

[33] Fabrice Delice, Moira Rousseau, and Jennifer Feitosa. 2019. Advancing teams research: What, when, and how to measure team dynamics over time. *Frontiers in psychology* 10 (2019), 1324. https://doi.org/10.3389/fpsyg.2019.01324

[34] Andy Dong. 2004. Quantifying coherent thinking in design: a computational linguistics approach. In *Design computing and cognition'04*. Springer, 521–540. https://doi.org/10.1007/978-1-4020-2393-4_27

[35] Andy Dong. 2005. The latent semantic approach to studying design team communication. *Design studies* 26, 5 (2005), 445–461. https://doi.org/10.1016/j.destud.2004.10.003

[36] Andy Dong, Andrew W Hill, and Alice M Agogino. 2004. A document analysis method for characterizing design team performance. *J. Mech. Des.* 126, 3 (2004), 378–385. https://doi.org/10.1115/1.1711818

[37] Debbie Dougherty and Kristina Drumheller. 2006. Sensemaking and Emotions in Organizations: Accounting for Emotions in a Rational(ized) Context. *Communication Studies* 57 (07 2006), 215–238. https://doi.org/10.1080/10510970600667030

[38] Melanie Duckert, Louise Barkhuus, and Pernille Bjørn. 2023. Collocated distance: a fundamental challenge for the design of hybrid work technologies. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16. https://doi.org/10.1145/3544548.3580899

[39] Tom Duffy, Chris Baber, and Neville A Stanton. 2013. Measuring collaborative sensemaking.. In *Iscram*.

[40] Casey Dugan, Werner Geyer, Michael Muller, Abel N Valente, Katherine James, Steve Levy, Li-Te Cheng, Elizabeth Daly, and Beth Brownholtz. 2012. "I'd never get out of this !?$%# office" redesigning time management for the enterprise. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1755–1764. https://doi.org/10.1145/2207676.2208307

[41] Amy Edmondson. 1999. Psychological safety and learning behavior in work teams. *Administrative science quarterly* 44, 2 (1999), 350–383. https://doi.org/10.2307/2666999

[42] Elliot E Entin and Daniel Serfaty. 1999. Adaptive team coordination. *Human factors* 41, 2 (1999), 312–325. https://doi.org/10.1518/001872099779591196

[43] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5, 1 (March 2006), 80–92. https://doi.org/10.1177/160940690600500107

[44] Sharon Ferguson, Kimberly Lai, James Chen, Safa Faidi, Kevin Leonardo, and Alison Olechowski. 2022. "Why couldn't we do this more often?": exploring the feasibility of virtual and distributed work in product design engineering. *Research in engineering design* 33, 4 (2022), 413–436. https://doi.org/10.1007/s00163-022-00391-2

[45] Sharon A Ferguson, Kathy Cheng, Lauren Adolphe, Georgia Van de Zande, David Wallace, and Alison Olechowski. 2022. Communication patterns in engineering enterprise social networks: an exploratory analysis using short text topic modelling. *Design Science* 8 (2022), e18. https://doi.org/doi:10.1017/dsj.2022.12

[46] Sharon A Ferguson, Georgia Van de Zande, and Alison Olechowski. 2024. No Risk, No Reward: Towards An Automated Measure of Psychological Safety from Online Communication. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7. https://doi.org/10.1145/3613905.3650923

[47] Clayton Feustel, Shyamak Aggarwal, Bongshin Lee, and Lauren Wilcox. 2018. People like me: Designing for reflection on aggregate cohort data in personal informatics systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–21. https://doi.org/10.1145/3264917

[48] Gerhard Fischer and Jonathan Ostwald. 2005. Knowledge communication in design communities. In *Barriers and biases in computer-mediated knowledge communication: And how they may be overcome*. Springer, 213–242. https://doi.org/10.1007/0-387-24319-4_10

[49] David Fono and Ron Baecker. 2006. Structuring and supporting persistent chat conversations. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. 455–458. https://doi.org/10.1145/1180875.1180944

[50] Heather Friedberg, Diane Litman, and Susannah B. F. Paletz. 2012. Lexical entrainment and success in student engineering groups. In *2012 IEEE Spoken Language Technology Workshop (SLT)*. 404–409. https://doi.org/10.1109/SLT.2012.6424258

[51] Daniel Y Fu, Simran Arora, Jessica Grogan, Isys Johnson, Sabri Eyuboglu, Armin W Thomas, Benjamin Spector, Michael Poli, Atri Rudra, and Christopher Ré. 2023. Monarch Mixer: A Simple Sub-Quadratic GEMM-Based Architecture. In *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.2310.12109

[52] Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research* 37, 1 (2010), 3–19. https://doi.org/10.1177/0093650209351468

[53] Sandy JJ Gould. 2024. Differential privacy and collective bargaining over workplace data. *Italian Labour Law e-Journal* 17, 2 (2024), 133–144. https://doi.org/10.6092/issn.1561-8048/20838

[54] Sandy J. J. Gould. 2024. Stochastic Machine Witnesses at Work: Today's Critiques of Taylorism are Inadequate for Workplace Surveillance Epistemologies of the Future. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3613904.3642206

[55] Sandy J. J. Gould, Anna Rudnicka, Dave Cook, Marta E. Cecchinato, Joseph W. Newbold, and Anna L. Cox. 2023. Remote Work, Work Measurement and the State of Work Research in Human-Centred Computing. *Interacting with Computers* 35, 5 (Dec. 2023), 725–734. https://doi.org/10.1093/iwc/iwad014

[56] Steve Hatfield, Brad Kreit, and Cantrell. [n. d.]. *Navigating the data dilemma: Can organizations build trust while using workforce data to improve performance?* Technical Report. https://www.deloitte.com/us/en/insights/topics/talent/monitoring-employees-in-the-workplace.html

[57] Jeffrey Heer and Maneesh Agrawala. 2008. Design considerations for collaborative visual analytics. *Information visualization* 7, 1 (2008), 49–62. https:

//doi.org/10.1057/palgrave.ivs.9500167
[58] Andrew Hill, Shuang Song, Andy Dong, and Alice Agogino. 2001. Identifying shared understanding in design using document analysis. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 80258. American Society of Mechanical Engineers, 309–315. https://doi.org/10.1115/DETC2001/DTM-21713
[59] Pamela J Hinds and Suzanne P Weisband. 2003. Knowledge sharing and shared understanding in virtual teams. *Virtual teams that work: Creating conditions for virtual team effectiveness* (2003), 21–36.
[60] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24. https://doi.org/10.1145/642611.642616
[61] Molly E Ireland and Marlone D Henderson. 2014. Language style matching, engagement, and impasse in negotiations. *Negotiation and conflict management research* 7, 1 (2014), 1–16. https://doi.org/10.1111/ncmr.12025
[62] Molly E Ireland and James W Pennebaker. 2010. Language style matching in writing: synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology* 99, 3 (2010), 549.
[63] Molly E Ireland, Richard B Slatcher, Paul W Eastwick, Lauren E Scissors, Eli J Finkel, and James W Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological science* 22, 1 (2011), 39–44. https://doi.org/10.1177/0956797610392928
[64] Petra Isenberg, Niklas Elmqvist, Jean Scholtz, Daniel Cernea, Kwan-Liu Ma, and Hans Hagen. 2011. Collaborative visualization: Definition, challenges, and research agenda. *Information Visualization* 10, 4 (2011), 310–326. https://doi.org/10.1177/1473871611412817
[65] Gemma Jiang. 2023. Collaborative leadership in team science: dynamics of sense making, decision making, and action taking. *Frontiers in Research Metrics and Analytics* 8 (2023), 1211407. https://doi.org/10.3389/frma.2023.1211407
[66] Tristan E Johnson and Debra L O'Connor. 2008. Measuring team shared understanding using the analysis-constructed shared mental model methodology. *Performance Improvement Quarterly* 21, 3 (2008), 113–134. https://doi.org/10.1002/piq.20034
[67] Jon R Katzenbach and Douglas K Smith. 2005. The discipline of teams. *Harvard business review* 83, 7 (2005), 162. https://hbr.org/1993/03/the-discipline-of-teams-2
[68] Harmanpreet Kaur, Daniel McDuff, Alex C Williams, Jaime Teevan, and Shamsi T Iqbal. 2022. "I didn't know I looked angry": Characterizing observed emotion and reported affect at work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18. https://doi.org/10.1145/3491102.3517453
[69] Anna Kawakami, Shreya Chowdhary, Shamsi T Iqbal, Q Vera Liao, Alexandra Olteanu, Jina Suh, and Koustuv Saha. 2023. Sensing wellbeing in the workplace, why and for whom? envisioning impacts with organizational stakeholders. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–33. https://doi.org/10.1145/3610207
[70] Ryan Kelly. 2015. Pyenchant. https://pypi.org/project/pyenchant/
[71] Katherine J. Klein and Steve W. J. Kozlowski. 2000. *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions*. Wiley.
[72] Rajesh Kumar and Paul Wilson. 2023. The Role of Data Visualization in Enhancing Communication between Technical and Non-Technical Teams. *Available at SSRN 5798462* (2023). https://doi.org/10.2139/ssrn.5798462
[73] Carine Lallemand, Alina Lushnikova, Joshua Dawson, and Romain Toebosch. 2024. Trinity: A Design Fiction to Unravel the Present and Future Tensions in Professional Informatics and Awareness Support Tools. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*. 1–15. https://doi.org/10.1145/3663384.3663396
[74] Janice Langan-Fox. 2003. Skill acquisition and the development of a team mental model: an integrative approach to analysing organizational teams, task, and context. *International Handbook of Organizational Teamwork and Cooperative Working* (2003), 321–360. https://doi.org/10.1002/9780470696712.ch16
[75] Janice Langan-Fox, Jeromy Anglim, and John R Wilson. 2004. Mental models, team mental models, and performance: Process, development, and future directions. *Human Factors and Ergonomics in Manufacturing & Service Industries* 14, 4 (2004), 331–352. https://doi.org/10.1002/hfm.20004
[76] Gilly Leshed, Jeffrey T Hancock, Dan Cosley, Poppy L McLeod, and Geri Gay. 2007. Feedback for guiding reflection on teamwork practices. In *Proceedings of the 2007 ACM International Conference on Supporting Group Work*. 217–220. https://doi.org/10.1145/1316624.1316655
[77] Laurie L Levesque, Jeanne M Wilson, and Douglas R Wholey. 2001. Cognitive divergence and shared mental models in software development project teams. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 22, 2 (2001), 135–144. https://doi.org/10.1002/job.87
[78] Katharina Lix, Amir Goldberg, Sameer B Srivastava, and Melissa A Valentine. 2022. Aligning differences: Discursive diversity and team performance. *Management Science* 68, 11 (2022), 8430–8448. https://doi.org/10.1287/mnsc.2021.4274
[79] Alina Lushnikova, Kerstin Bongard-Blanchy, and Carine Lallemand. 2022. What aspects of collaboration are meaningful to you? informing the design of self-tracking technologies for collaboration. In *Adjunct Proceedings of the 2022 Nordic Human-Computer Interaction Conference*. 1–5. https://doi.org/10.1145/3547522.3547681
[80] Sally Maitlis and Marlys Christianson. 2014. Sensemaking in Organizations: Taking Stock and Moving Forward. *The Academy of Management Annals* 8 (06 2014). https://doi.org/10.1080/19416520.2014.873177
[81] Wendy Martinez, Johann Benerradi, Serena Midha, Horia A Maior, and Max L Wilson. 2022. Understanding the ethical concerns for neurotechnology in the future of work. In *Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work*. 1–19. https://doi.org/10.1145/3533406.3533423
[82] Sara McComb, Deanna Kennedy, Rebecca Perryman, Norman Warner, and Michael Letsky. 2010. Temporal patterns of mental model convergence: Implications for distributed teams interacting in electronic collaboration spaces. *Human Factors* 52, 2 (2010), 264–281. https://doi.org/10.1177/0018720810370458
[83] Sara A McComb. 2007. Mental model convergence: The shift from being an individual to being a team member. In *Multi-level issues in organizations and time*. Vol. 6. Emerald Group Publishing Limited, 95–147. https://doi.org/Mentalmodelconvergence:Theshiftfrombeinganindividualtobeingateammember.
[84] Susan Mohammed, Lori Ferzandi, and Katherine Hamilton. 2010. Metaphor no more: A 15-year review of the team mental model construct. *Journal of management* 36, 4 (2010), 876–910. https://doi.org/10.1177/0149206309356804
[85] Lillio Mok, Lu Sun, Shilad Sen, and Bahareh Sarrafzadeh. 2023. Challenging but Connective: Large-Scale Characteristics of Synchronous Collaboration Across Time Zones. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17. https://doi.org/10.1145/3544548.3581141
[86] Michele Molè. 2024. Commodified, Outsourced Authority: A Research Agenda for Algorithmic Management at Work. *Italian Labour Law e-Journal* 17, 2 (Dec. 2024), 169–188. https://doi.org/10.6092/issn.1561-8048/20836
[87] Phoebe Moore, Lukasz Piwek, and Ian Roper. 2018. The quantified workplace: A study in self-tracking, agility and change management. *Self-Tracking: Empirical and philosophical investigations* (2018), 93–110. https://doi.org/10.1007/978-3-319-65379-2_7
[88] Mimi Nguyen, Milad Laly, Bum Chul Kwon, Celine Mougenot, and John McNamara. 2022. Moody Man: Improving creative teamwork through dynamic affective recognition. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–14. https://doi.org/10.1145/3491101.3519656
[89] Laura Okpara, Colin Werner, Adam Murray, and Daniela Damian. 2023. The role of informal communication in building shared understanding of non-functional requirements in remote continuous software engineering. *Requirements Engineering* 28, 4 (2023), 595–617. https://doi.org/10.1007/s00766-023-00404-z
[90] James W Pennebaker and Cindy K Chung. 2012. Language and social dynamics. In *Technical Report 1318*. University of Texas at Austin.
[91] Erika Peterson, Terence R. Mitchell, Leigh Thompson, and Renu Burr. 2000. Collective Efficacy and Aspects of Shared Mental Models as Predictors of Performance Over Time in Work Groups. *Group Processes & Intergroup Relations* 3, 3 (July 2000), 296–316. https://doi.org/10.1177/1368430200033005
[92] Martin J Pickering and Simon Garrod. 2004. The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences* 27, 2 (2004), 212–225. https://doi.org/10.1017/s0140525x04450055
[93] Kalle A Piirainen, Gwendolyn L Kolfschoten, and Stephan Lukosch. 2012. The joint struggle of complex engineering: A study of the challenges of collaborative design. *International Journal of Information Technology & Decision Making* 11, 06 (2012), 1087–1125. https://doi.org/10.1142/S0219622012400160
[94] Ray E Reagans, Hagay Volvovsky, and Ronald S Burt. 2023. Shared language in the team network-performance association: Reconciling conflicting views of the network centralization effect on team performance. *Collective Intelligence* 2, 3 (2023), 26339137231199739. https://doi.org/10.1177/26339137231199739
[95] Receptiviti Inc. 2022. receptiviti: Text Analysis Through the Receptiviti API. https://receptiviti.github.io/receptiviti-r/
[96] Eduardo Ed Salas and Stephen M Fiore. 2004. *Team cognition: Understanding the factors that drive process and performance*. American Psychological Association. https://doi.org/10.1037/10690-000
[97] Samiha Samrose, Daniel McDuff, Robert Sim, Jina Suh, Kael Rowan, Javier Hernandez, Sean Rintel, Kevin Moynihan, and Mary Czerwinski. 2021. Meetingcoach: An intelligent dashboard for supporting effective & inclusive meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3411764.3445615
[98] Samiha Samrose, Ru Zhao, Jeffery White, Vivian Li, Luis Nova, Yichen Lu, Mohammad Rafayet Ali, and Mohammed Ehsan Hoque. 2018. Coco: Collaboration coach for understanding team dynamics during video conferencing. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 4 (2018),

1–24. https://doi.org/10.1145/3161186
[99] Catarina Marques Santos, Sjir Uitdewilligen, and Ana Margarida Passos. 2015. Why is your team more creative than mine? The influence of shared mental models on intra-group conflict, team creativity and effectiveness. *Creativity and innovation management* 24, 4 (2015), 645–658. https://doi.org/10.1111/caim.12129
[100] Beau G Schelble, Christopher Flathmann, Nathan J McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's think together! Assessing shared mental models, performance, and trust in human-agent teams. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–29. https://doi.org/10.1145/3492832
[101] Shuang Song, Andy Dong, Alice M Agogino, et al. 2003. Time variation of design "story telling" in engineering design teams. In *DS 31: Proceedings of ICED 03, the 14th International Conference on Engineering Design, Stockholm*. 423–424.
[102] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics* 26, 3 (2000), 339–374. https://aclanthology.org/J00-3003/
[103] Jina Suh, Javier Hernandez, Koustuv Saha, Kathy Dixon, Mehrab Bin Morshed, Esther Howe, Anna Kawakami, and Mary Czerwinski. 2023. Towards successful deployment of wellbeing sensing technologies: Identifying misalignments across contextual boundaries. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 1–8. https://doi.org/10.1109/ACIIW59127.2023.10388088
[104] Petrus Te Braak, Theun Pieter van Tienoven, Joeri Minnen, and Ignace Glorieux. 2023. Data quality and recall bias in time-diary research: The effects of prolonged recall periods in self-administered online time-use surveys. *Sociological Methodology* 53, 1 (2023), 115–138. https://doi.org/10.1177/00811750221126499
[105] The Microsoft 365 Marketing Team. 2018. New survey explores the changing landscape of teamwork. https://www.microsoft.com/en-us/microsoft-365/blog/2018/04/19/new-survey-explores-the-changing-landscape-of-teamwork/
[106] Together AI. 2024. Together AI. https://together.ai
[107] Roy Van Den Heuvel and Carine Lallemand. 2023. Personal Informatics at the Office: User-Driven, Situated Sensor Kits in the Workplace. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*. 1–13. https://doi.org/10.1145/3596671.3598577
[108] Elizabeth S Veinott, Judith Olson, Gary M Olson, and Xiaolan Fu. 1999. Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 302–309. https://doi.org/10.1145/302979.303067
[109] Radim vRehuuvrek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. (2010).
[110] Karl E Weick, Kathleen M Sutcliffe, and David Obstfeld. 2005. Organizing and the process of sensemaking. *Organization science* 16, 4 (2005), 409–421. https://doi.org/10.1287/orsc.1050.0133
[111] Jaime Windeler, Likoebe M Maruping, Lionel P Robert, and Cynthia K Riemenschneider. 2015. E-profiles, conflict, and shared understanding in distributed teams. *Journal of the Association for Information Systems* 16, 7 (2015), 1. https://doi.org/10.17705/1jais.00401
[112] Y Jasmine Wu, Brennan Antone, Leslie DeChurch, and Noshir Contractor. 2023. Information sharing in a hybrid workplace: understanding the role of ease-of-use perceptions of communication technologies in advice-seeking relationship maintenance. *Journal of Computer-Mediated Communication* 28, 4 (June 2023), zmad025. https://doi.org/10.1093/jcmc/zmad025
[113] Longqi Yang, David Holtz, Sonia Jaffe, Siddharth Suri, Shilpi Sinha, Jeffrey Weston, Connor Joyce, Neha Shah, Kevin Sherman, Brent Hecht, and Jaime Teevan. 2021. The effects of remote work on collaboration among information workers. *Nature Human Behaviour* (2021). https://doi.org/10.1038/s41562-021-01196-4
[114] Xi Yang, Martin Helander, Andy Dong, et al. 2009. Latent semantic analysis measures participant affiliation and team process quality. In *DS 58-9: Proceedings of ICED 09, the 17th International Conference on Engineering Design, Vol. 9, Human Behavior in Design, Palo Alto, CA, USA, 24.-27.08. 2009*. 35–46.
[115] Sue Yi, Nicole B Damen, and Christine A Toh. 2020. Back and forth: Using conversation analysis to explore dialogues of sharedness of mental models. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 83976. American Society of Mechanical Engineers, V008T08A025. https://doi.org/10.1115/DETC2020-22515

A SHARED UNDERSTANDING METRICS

Language Style Matching: In Equations 1 and 2, c_a, c_b are frequency measures for linguistic category c for team members a and b, C is the total number of linguistic categories being studied, and n is the number of consenting team members who sent a message on the chosen channel on the chosen day. $\text{LSM}_{\text{group}}$ is then displayed over time (Figure 3).

$$\text{LSM}_{\text{category}} = 1 - \frac{|c_a - c_b|}{c_a + c_b + 0.001} \tag{1}$$

$$\text{LSM}_{\text{group}} = \sum_{a=1}^{n} \sum_{\substack{b=1 \\ a \neq b}}^{n} \left(\frac{1}{C} \sum_{c=1}^{C} \text{LSM}_{\text{category}}(c_a, c_b) \right) \tag{2}$$

Knowledge Convergence: The aggregated messages are tokenized and stop words are removed. Then, a document-term matrix is created that describes how commonly each term is used in each document, where a document is the aggregation of all messages sent by a user in a channel on a day. Using the Gensim package [109], LSA was applied to the document-term matrix. To choose the appropriate number of topics (or the k vectors to retain in the resulting LSA matrix), we iterate through topics from 2 to $number of documents/5$, recording the c_v coherence measure from the Gensim package, and the model with the highest coherence is retained. Using this model, for each channel and day, every team member's messages are transformed into a vector, where the group value for that day is calculated as the centroid of all team members' vectors. The "total knowledge" defined for that time period is the group centroid on the *last day of the time period*. Then, for each day, we calculate the cosine similarity between each team member's centroid, the team average centroid, and the "total knowledge" centroid. Thus, this plot describes how quickly the team converges on their knowledge, and who within the team is driving this convergence (Figure 3).

Semantic Coherence: The initial steps in the algorithm are the same as for Knowledge Convergence: tokenized messages are turned into a document-term matrix, input into LSA, where the number of topics is chosen by optimizing for c_v values. Semantic coherence is a group-level value, calculated by taking the sum of each team member's LSA-resultant vector for each day, and then taking the norm of this vector [35] (Equation 3 where m represents number of consenting team members). Since each team member's vector represents the distribution of "topics" discussed, if all team members had similar distributions, the norm of the vector would be higher. We scale this value to be between 0-1 for consistency with the other plots (Figure 3).

$$\chi = \left\| \sum_{m} s_i \right\| \tag{3}$$

Embedding Space Similarity: This value is calculated similarly to LSM where the embedding space vectors for each possible pair within the team are compared, and the similarity is averaged across the whole group. We use the Together AI API to access embedding models [106] as it has the strongest privacy measures, and we use the 'm2-bert-80M-32k-retrieval' model to capture message embeddings [51]. For each channel, day, and user, we input the aggregated messages into the model to retrieve an embedding vector. We then calculate the cosine similarity between each possible pair in the team for each day, and average this similarity across all pairs to get an average vector for the team on each day, which is then plotted for the user (Figure 3).

B DATA FLOW DIAGRAM

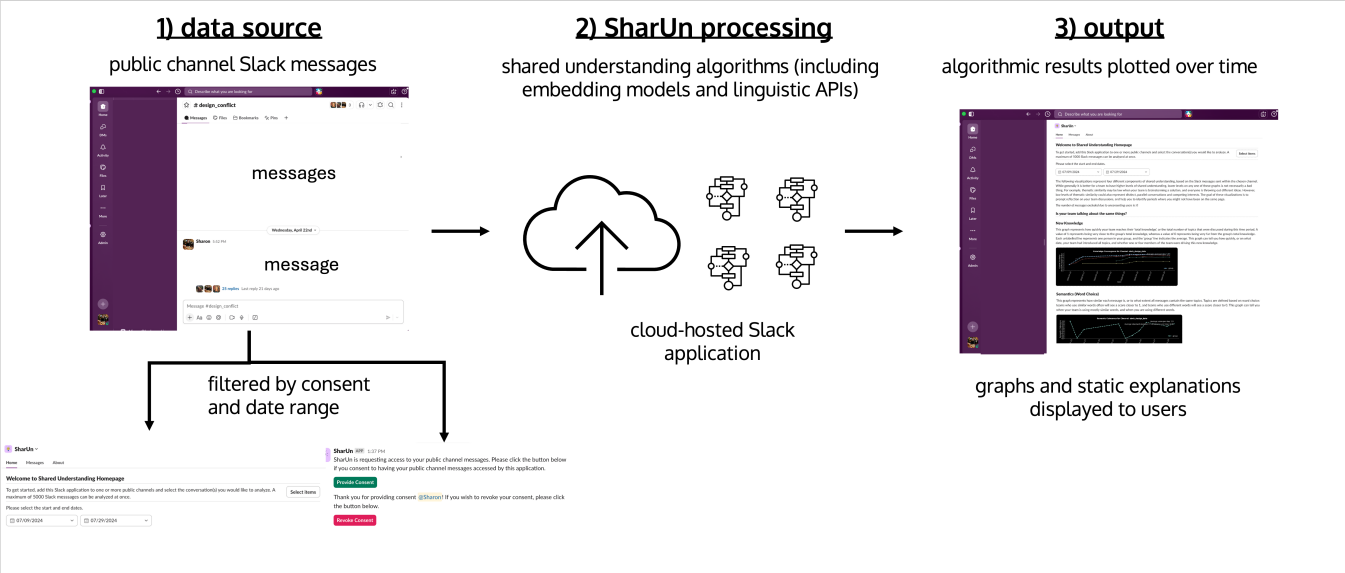

Figure 8: Data flow for the SharUn technology probe. Public channel Slack messages were filtered by the selected date range and only those message sent by consenting team members were retained. This data was sent to the AWS-hosted Slack application, where APIs were used in the processing of the four shared understanding algorithms. Then, these four algorithms were plotted over time, and these images were sent back to the Slack application, and displayed alongside static explanations for the user.

C QUANTITATIVE STUDY - SURVEY QUESTIONS

Table 3: Survey questions

Construct	Question
Shared Goals	In this team, we share a common vision.
	In this team, we act toward common goals
	Members of this team act without having a clear direction (reverse-scored)
Shared Knowledge	Members of this team know what tasks other team members deal with
	In this team, we share with one another the tasks we work on
	Sharing with one another in the context of the course project enables us to better understand the needs of each other
	Sharing with one another about our course project issues enables us to better understand how our actions impact other team members

D SURVEY STANDARD DEVIATION

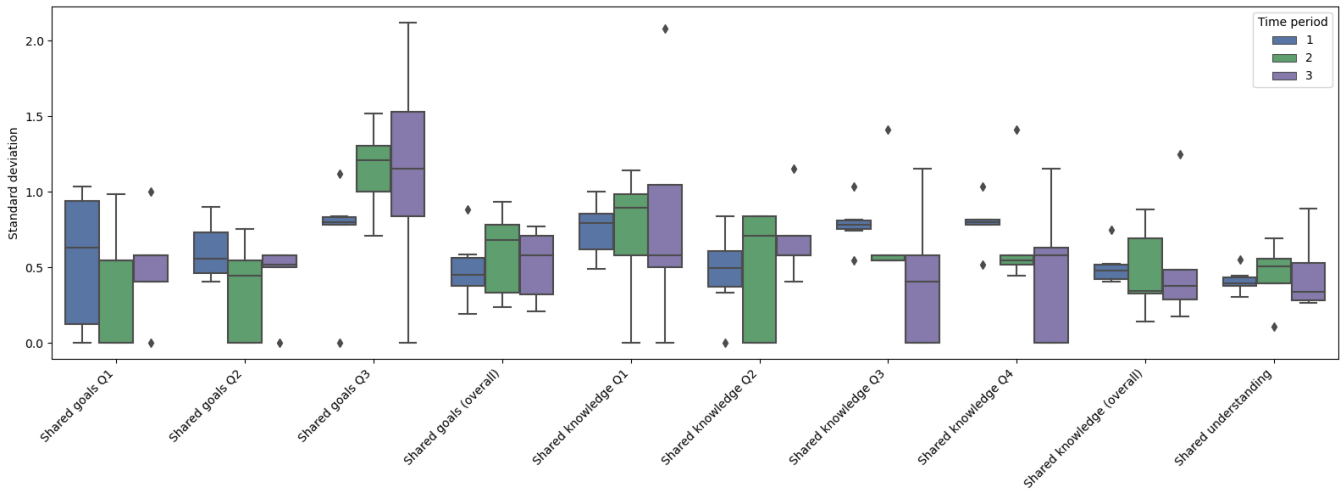

Figure 9: The full range of per-team standard deviations for each question in the measure of shared understanding, per time period.

Received 3 April 2026; accepted 22 May 2026