Probabilistic Modeling of Interpersonal Coordination Processes

Paulo Soares [1] **Adarsh Pyarelal** [1] **Meghavarshini Krishnaswamy** [1] **Emily Butler** [1] **Kobus Barnard** [1]

Abstract

We develop a novel probabilistic model for interpersonal coordination as a latent phenomenon explaining statistical temporal influence between multiple components in a system. For example, the state of one person can influence that of another at a later time, as indicated by their observed behaviors. We characterize coordination as the degree to which the distributions for such states at one time point are merged for the next salient time point. We evaluate our model in the context of three-person teams executing a virtual search and rescue (SAR) mission. We first use synthetic data to confirm that our technical definition of coordination is consistent with expectations and that we can recover generated coordination despite noise. We then show that captured coordination can be predictive of team performance on real data. Here we use speech vocalics and semantics to infer coordination for 36 teams carrying out two successive SAR missions. In two different datasets, we find that coordination is generally predictive of team score for the second mission, but not for the first, where teams are largely learning to play the game. In addition, we found that including a semantic modality improves prediction in some scenarios. This shows that our intuitive technical definition can capture useful explanatory aspects of team behavior.

1. Introduction

In a basketball game, when a team is on the offense, they must maintain proper spacing and timing to execute plays effectively without interfering with each other's positions. When on defense, they must work together to switch assignments, cover open areas, and block shots. In other words, they need to be *coordinated*. This example illustrates that the concept of *coordination* is intuitively understood by most people, i.e., 'you know it when you see it'. However, defining it precisely—and computationally—is challenging.

We study coordination in the context of a system of humans working together towards a common goal (i.e., collaboration) or contrastive goals (i.e., competition). Coordination here means the spontaneous temporal synchronization of behavioral sub-processes among individuals in social interactions, known as interpersonal coordination (Cornejo et al., 2017; Butler, 2022). Computationally, we interpret coordination as the causal linkage of a latent process that captures important interactions and could predict outcomes. This perspective on coordination does not require agency, goal-directed behavior, or modeling as a game, unlike the game-theoretic view (Cooper, 1999). It focuses on how one system abstraction (e.g., communication behavior), influences another.

Such abstraction is needed to endow systems with interpretable understanding of complex interactions among humans. We are motivated to assess coordination in scenarios where it remains implicit, subliminal, unconscious, or perhaps salient to participants but not easily explained, such as when two individuals seamlessly connect while performing a task. We are specifically interested in temporal correlations resulting from *causal interactions* between system components, excluding coordination due to unrelated factors—e.g., correlated brain activity between people watching the same movie simultaneously, without being in the same place (Hasson et al., 2004). This phenomenon is often labeled as coordination, but we prefer the term 'synchrony'. It is typically assessed through data correlations over time, but these correlations do not reflect interpersonal interaction; rather, they are tied to shared external factors, such as the movie being co-watched in this example.

Most current approaches do not directly address coordination front and center as an explanatory concept. It is commonly assumed that temporal correlations reflect coordination without defining what it means computationally. This limits the interpretability of coordination quantities and makes integrating signals from different modalities and timescales complex.

Notably, Butner et al. (2014)—influenced by von Holst

[*] Equal contribution [1] University of Arizona. Correspondence to: Paulo Soares <paulosoares@arizona.edu>, Adarsh Pyarelal <adarsh@arizona.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

(1973)—offers a broader approach where change scores in two variables are driven by a common coordination level, but it does not consider causal influence between interacting components. Recently, Wiltshire et al. (2022) posited the pressing need for a more systematic approach to coordination, especially in online contexts. Inspired by this urgency and the admirable prior efforts to quantify coordination thus far, we introduce a novel computational definition to overcome the limitations of existing methods.

Key Contributions 1) We propose a novel computational model that explicitly defines coordination as a latent phenomenon that controls the influence of multiple components in a system on each other at a later point in time; **2)** We develop how evidence for coordination processes can be from diverse modalities, operating at different time scales. **3)** Using synthetic data, we propose a measure to evaluate when estimating coordination is useful and show that our computational definition of coordination aligns with our intuition and can handle injected noise in the data. **4)** Using data from real human participants, we show that our model of coordination can be predictive of team performance on collaborative tasks and that adding extra modalities can be advantageous.

2. Related Work

Most existing approaches address coordination [1] using time-series analysis techniques to identify temporal correlations. These proxy measures for coordination are then studied as informative of group attributes such as team performance. Cross-correlation of brain signals from interacting partners has been found to be predictive of team performance (Henning & Korbelak, 2005). Entropy yielded a similar trend in experiments with dialog data (Wiltshire et al., 2018; Engome Tchupo & Macht, 2023; Delaherche et al., 2012). Some studies (Levitan & Hirschberg, 2011; Lubold & Pon-Barry, 2014; Borrie et al., 2015; Litman et al., 2016; Rahimi et al., 2017; 2019) quantify coordination by comparing the values of multiple time series in terms of similarity and convergence. More abstractly, principal component analysis can assess interpersonal coordination as the degree to which variance is preserved when one person's series is projected onto the space of another person's series (Lee et al., 2014).

Cross-recurrence quantification analysis (CRQA) (Marwan et al., 2002; Wallot, 2019) has been used to quantify aspects of coordination and team dynamics (Strang et al., 2014; Fusaroli et al., 2016; Knight et al., 2016; Borrie et al., 2019; Amon et al., 2019; Borrie et al., 2019). Similar analyses have been reported in the frequency domain using tools like cross-wavelet coherence (CWC) to estimate coordination in collaborative problem-solving (Wiltshire et al., 2019), jazz performances (Walton et al., 2015), joke-telling (Schmidt et al., 2012), unstructured conversations (Fujiwara & Daibo, 2016), and dance (Washburn et al., 2014).

Mechanistic Models Instead of focusing on temporal data correlations, another approach consists of modeling coordination using a set equations defining a system of coupled oscillators (Zhang et al., 2019; Miao et al., 2023). For instance, Zhang et al. (2019) define coordination as a phenomenon influencing coupling strength through phase-locking. The authors found success in capturing coordination patterns with and across groups in a multiparty rhythmic task. Rather than define coordination as synchronous phase-locking, we define it as a level of causal influence among latent processes, themselves linked to varying data modalities and time scales.

Latent Variable Models More similar to our work, Butner et al. (2014) employ latent variables to capture coordination in multimodal data streams over time, focusing on synchrony control among modalities but not causality. Their reliance on Latent Change Score (LCS) models presents challenges such as complex interpretation and implementation, limitations in handling non-linear relationships, and biased estimation (Klopack & Wickrama, 2019; Kievit et al., 2018; Fernández et al., 2021). In a similar vein, Moulder et al. (2022) introduce a framework to capture temporal dependencies and interactions between multiple data modalities, but do not explicitly define coordination as a construct. Instead, they rely on a synchrony measure, aligning more with Butner et al.'s definition. In contrast to these two approaches, we adopt dynamic Bayes networks (DBNs) due to their generative capacity, flexibility in incorporating prior knowledge, and ability to handle non-linear relationships, missing data, and uncertainty (Murphy, 2012).

Multi-agent Reinforcement Learning Our view on coordination differs from multi-agent reinforcement learning (MARL), but there is some overlap in situations where humans synchronize efforts toward shared or complementary objectives. For example, in the Minecraft-based search and rescue missions we study in this paper, we sometimes assume that the participants' shared goal is a higher team score. However, our approach is agnostic to this goal-oriented explanation of human behavior. For instance, the players could equally want to socially connect with their peers. More generally, interpersonal coordination can encompass scenarios with implicit, unknown or even non-existent goals (e.g., a couple arguing because there are unresolved power issues in their relationship that they cannot articulate). Further, in our applications, the notion of local utility would be more abstract, such as the utility of a change in neural activity or respiratory rhythm in resolving the ill-defined issue with the

[1] The term 'entrainment' is used in some of these studies to refer to a concept that is qualitatively similar to our working definition of coordination.

couple's power dynamics.

A common limitation among the current approaches is the difficulty of combining data with different time scales in the model. This is important because different modalities may be observed at different frequencies that define not only the periodicity of raw measurements but also how they affect each other in the system. In the next section, we detail how our formulation naturally combines components with different time scales.

3. Model of Coordination

Intuitively, coordination is a general concept present in many contexts, across multiple observation modalities and time scales. We observe coordination among people in groups as a phenomenon detached from particular observation modalities. Hence, a model of coordination should be about the statistical properties of the interacting components. We execute this by having coordination explicitly modulate the degree to which components are influencing each other. We further model components of the system as latent so that their temporal trajectory can simply lead to observations of multiple modalities at arbitrary time scales.

Main Idea We model a coordination process as some latent components being repeatedly predictive of others at some *later* point in time. Consider two components, each with a distribution for their next state (e.g., a normal distribution centered at the current state). If each component draws from its own distribution, then the components are not influencing each other. A key idea of our approach is to identify coordination as using some of one person's distribution to predict another's distribution. This excludes spurious correlations such as the example in §1 of two people watching the same movie in different places. Here, your own state might predict your next state due to smoothness, but using the other person's predicted state would generally not be additionally helpful. However, if they are sitting next to you and begin laughing at the movie, causing you to also laugh, then their state has some predictive power for your next state and we consider the phenomenon as coordination in our model.

Latent System Components We denote latent system components that may exhibit collective coordination by $A_t^{n,p}$, where n indexes components (e.g., speech vocalics), p indexes people, and t indexes discrete time steps associated with available data. We denote the collection of components by $\mathcal{A} = \{A^{n,p}\}$. The time scale for each A^n can be different—hence, the semantics of the subscript t depend on the context implied by A^n. When this needs to be explicit, we use t_n and map to actual time by $f^n(t_n)$. For instance, if we model coordination at every second but observe some modality, n, at regular intervals of 5 seconds, $f^n(t^n) = 5t^n$.

The precise identities of these latent components remain an empirical question. However, we have an intuition on what they may represent. For instance, cognitive state is a possible interpretation for a latent system component modeling brain data. Alternative possibilities are attentional state, emotional state, or mental effort.

Component Groups In our latent space, the level of coordination C is shared by multiple interacting groups of components, $I_g = \{A^{n_m^g, p_m^g}\}_{m=1:M_g}$, where g indexes the groups, m indexes the members of a group, and M_g is the number of members of group I_g.

In other words, component groups are used to group modalities into one latent system component. Typically, these groups exhibit similar time scales and consist of a single evidence modality; however, this criterion is not mandatory. For instance, it is possible to employ a single group to encapsulate latent components that depict neural activity as observed through functional near-infrared spectroscopy (fNIRS) and electroencephalogram (EEG) modalities simultaneously.

In this work we use two groups (latent vocalics and latent semantics) over three participant pairings.

Coordination Level We denote the *coordination level* by $C \in [0, 1]$. The latent components, $A^{n,p} \in \mathcal{A}$ within an interacting group influence each other as moderated by C, which controls how distributions for $A^{n,p}$ are blended. For a pair of components, $C = 0$ means that both are drawn from their own distribution, $C = 0.5$ means they share an equal blend of their distributions (maximally coordinated system), and $C = 1$ is the extreme case where they switch distributions.

In our experimental domain, coordination is symmetrically shared. If we were modeling a leader and follower, influence from coordination would be asymmetric. We also assume that coordination is relatively persistent—i.e., it is either constant, changes between 0 and 1 infrequently, or changes slowly, depending on the scenario. Figure 1 shows a graphical model with two interacting groups with different time scales: $I_1 = \{A^{1,1}, A^{1,2}\}$ and $I_2 = \{A^{2,1}, A^{2,2}\}$.

Observations We denote observations associated with $A_t^{n,p}$ as $O_t^{n,p,l} \in O$, where l indexes features, each with its own distribution.

Joint Distribution Let T be the number of time steps in the coordination time scale, G the number of interacting groups, $\mathcal{M} = \{M_1, ..., M_G\}$ the set containing the number of members in each interacting group, $T^{n_m^g}$ the number of time steps in the scale of the component n_m^g, and $L^{n_m^g}$ the number of features in observations of the component n_m^g. The joint distribution, $p(C_{0:T}, \mathcal{A}, \mathcal{O}, G, \mathcal{M})$, is the product of three terms, $p_{\text{coordination}}$, $p_{\text{component}}$, and $p_{\text{observation}}$:

$$\begin{aligned} p_{\text{coordination}} &= p\left(C_0\right) \prod_{t=1}^{T} p\left(C_t \mid C_{t-1}\right), \\ p_{\text{component}} &= \prod_{g=1}^{G} \prod_{m=1}^{M_g} p\left(A_0^{n_m^g, p_m^g}\right) p_{\text{transition}}, \qquad (1) \\ p_{\text{observation}} &= \prod_{g=1}^{G} \prod_{m=1}^{M_g} \prod_{t'=1}^{T_{n_m^g}} \prod_{l=1}^{L_{n_m^g}} p\left(O_{t'}^{n_m^g, p_m^g, l} \middle| A_{t'}^{n_m^g, p_m^g}\right), \end{aligned}$$

where $p_{\text{transition}}$ is given by

$$\prod_{t'=1}^{T^{n_m^g}} p\left(A_{t'}^{n_m^g, p_m^g} \middle| \left\{A_{t'-1}^{n_{m'}^g, p_{m'}^g}\right\}_{m'=1:M_g}, C_{f^{n_m^g}(t')}\right). \qquad (2)$$

While coordination does not appear in $p_{\text{observation}}$ directly, it may be convenient to have $p_{\text{observation}}$ as a function of coordination, which we do for semantic linkage in §6.

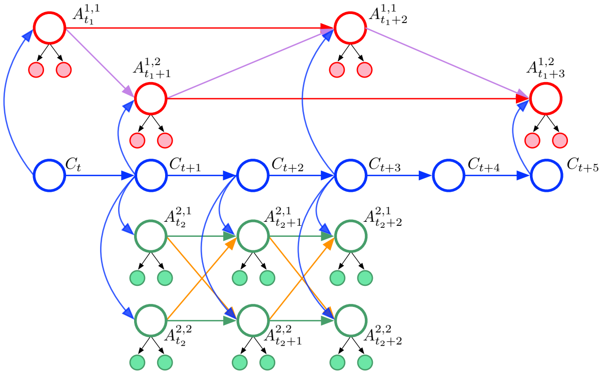

Figure 1: A dynamic Bayes net for coordination in the case of two pairs of latent conceptual components where each pair has its own time scale for linkage. Here, each component (red and green) is encoded for two people and has two observations, each at relevant times as indicated by shaded circles. As drawn, each person's component values influence themselves over time (red and green arrows) as regularization. The coordination variable (blue open circles) evolves over time and influences the degree that the distribution of a person's component is influenced by their partner's component at a previous time point. No coordination is equivalent to removing the lavender and orange arrows. For the discrete model, coordination simply selects between the two incoming distributions (e.g., red and lavender for $A^{1,1}$ and $A^{1,2}$). For the continuous model, the distributions are blended based on coordination.

3.1. Choices of Coordination Distribution

Discrete Coordination In our *discrete coordination* model, coordination is either present or absent—i.e., there are no intermediate levels. Formally, $C_t \in \{0, 1\}$ and is either declared constant ($C_t \equiv C$) or switches at each time step with some probability. For $c = C_{f^{n_m^g}(t')}$, the transition distribution in (2) becomes

$$p\left(A_{t'}^{n_m^g, p_m^g} \middle| A_{t'-1}^{n_m^g, p_m^g}\right)^{1-c} p\left(A_{t'}^{n_m^g, p_m^g} \middle| \left\{A_{t'-1}^{n_{m'}^g, p_{m'}^g}\right\}_{m \neq m'}\right)^{c}. \qquad (3)$$

That is, coordination defines whether a component depends on its base distribution ($c = 0$) or the base distribution for its peer(s) ($c = 1$). To explicitly model in-between behavior, we use *continuous coordination*.

Continuous Coordination For continuous coordination, instead of switching between the two distributions, we blend them. Executing this depends on the particular distributional and modeling intuition. For the simplest case of two normal distributions with means μ_0, μ_1 and equal variance σ_A^2, it is natural to use the convex combination of the two means, $\mu_c = (1-c)\mu_0 + c\mu_1$, as the mean for the blended distribution and σ_A^2 as its variance. The transition distribution from (2) becomes

$$\mathcal{N}\left(A_{t'}^{n_m^g, p_m^g} \middle| (1-c)A_{t'-1}^{n_m^g, p_m^g} + cA_{t'-1}^{n_{m'}^g, p_{m'}^g}, \sigma_A^2\right). \qquad (4)$$

3.2. Time Scales

Each interacting group of components shares a time scale for grouping associated observations that roughly corresponds to how quickly we expect distributional adaptation. For instance, brain activity from different regions might exhibit influence in milliseconds, while reacting to someone else's actions might occur in a few seconds. For vocalic entrainment (Lee et al., 2014)—a related phenomenon in which individuals mirror each other's speech and linguistic features during sustained interactions and cooperative tasks—a natural time scale is based on the speech utterance of one person delimited by turn-taking whereas for certain structured activities, such as team rope jumping, the activity itself imposes a clock on many observations. It is important not to confuse these time scales with the frequency of raw measurements. The coordination time scale is often implemented as windows of times where we define distributions of raw measurements. The modeling approach developed in this section naturally combines components with different time scales.

3.3. Going Beyond Pairs

We consider two cases for modeling components that are influenced by more than one other component. First, in cases like group conversations in our experiments in §7, we assume perfect turn-taking so that the influences are serialized (component $n = 1$ in Figure 1). Here we use metadata (e.g., the identity of the interlocutor) to dictate which variables participate in distributional blending at a given point in time, but we still have pairwise influences. The use of metadata here is minimal and straightforward. Implementation-wise, we provide components that abstract the use of metadata for model construction in our code.

A natural way to extend the pairwise strategy to N variables is to blend each variable's distribution with weight $1 - C$, with an equal blend (e.g., average of means) of the other $N - 1$ variables with weight C. So, for level $C = (N - 1)/N$, all variables have the same distribution (overall equal blend), for a maximally coordinated system. When $C > (N - 1)/N$ (we term this case *super-coordinated*), participants are mimicking other participants' distributions more than their own base distribution.

3.4. Model Instantiation

Our model of coordination is a general formulation about whether the distributions of your partners are helpful in predicting your own trajectory. We do not confine ourselves to the blending options proposed in this paper, nor do we take a definitive stance on whether coordination is discrete or continuous. Instead, we outline choices that we consider intuitive and easily applicable, serving as a foundational starting point for experimenters. Determining the most suitable blending scheme, conditional distributions, and coordination domain for a specific problem remains an empirical question.

Our findings with synthetic and real-world data suggest that opting for mean-based blending and continuous coordination proves advantageous. While we employ simple distributions in this study, we acknowledge the potential need for more complex ones. In such cases, neural networks may serve as optimal tools for learning intricate distributions, and our proposal does not preclude their usage.

Additionally, while our approach can have symmetric or unidirectional causality (e.g., a designated leader and follower), we chose to share the degree of causality among all subjects in our model instances (all players influence each other). This choice was largely in deference to our application, which is a collaborative task with no designated leaders and followers among the three participants.

4. Inference

The main contribution of our study lies in the proposed modeling approach. Appropriate learning methods will vary depending on the choices made when instantiating the conceptual framework into a precise model. In instances where a probabilistic graphical model (PGM) is employed, variants of sampling are likely to be effective at some computational cost. For more intricate model instances, gradient-based methods may prove to be more suitable.

In this work, we adopted NUTS (Hoffman & Gelman, 2014) as our inference method, as it demonstrated convergence in both synthetic and real-data evaluations and accurately estimated coordination in the synthetic model. Further details on configuration, and computational resources used can be found in Appendix A. Our code and preprocessed datasets are available online at https://github.com/ml4ai/tomcat-coordination.

5. Data

We evaluate our model on two distinct datasets based on the same task, but with different participants and experimental procedures. The first is the ASIST Study 3 dataset (Huang et al., 2022a), and the second is the ToMCAT dataset (Pyarelal et al., 2023). For the ASIST dataset, participants performed the task remotely from their own homes, while for the ToMCAT dataset, the participants were physically colocated in a lab and instrumented with a multitude of sensors including basic physiology (EKG), skin conductance (GSR), eye trackers, and combined fNIRS and EEG cap. In both cases, their spoken dialog during the missions was recorded. In this section, we offer a concise overview of the task and data cleaning process. For more details, refer to Appendix B.

Task In both experiments, teams played two missions, A and B, in which they were tasked with rescuing victims of an office building collapse simulated in Minecraft. The teammates were given unique roles with complementary abilities and pieces of information, incentivizing them to coordinate their activities and share information in order to maximize their performance on the task, as measured by a game score ranging from 0 to 950.

Potentially, teams were advised by one of six AI agents during the mission. From the ASIST dataset, we use data from two conditions: the 'No-Advisor' condition, where teams were not paired with an advisor (14 teams) and the 'ToMCAT Advisor' condition (13 teams). This particular advisor was designed to increase team coordination by intervening on team communications (Mathieu et al., 2020).

In the ToMCAT dataset, most of the trials used the ToMCAT advisor, and thus we only use data from that condition. After removing interrupted trials, trials with problems in the audio files for some participant, and trials where an experimenter had to stand-in as a third participant, we ended up with 16 trials (8 in each mission) from 9 teams.

Pre-processing From these datasets, we derived two modalities: *vocalic features* and *event labels*, extracted from spoken inter-player dialog transcribed by an automatic speech recognition (ASR) component. We analyze both modalities in units of utterances, assuming perfect turn-taking. In cases of overlapping utterances where multiple people speak simultaneously, we use both utterances as evidence, ordered by their end times. If their end times coincide, we randomly select one.

We use four openSMILE (Eyben et al., 2010) vocalic features (pitch, intensity, jitter, and shimmer) from each participant's audio stream. Using utterance start and end timestamps, we segment the stream to discard periods of silence. Averaging vocalic features within each utterance, we obtain a 4-dimensional sparse time series of average vocalic features per participant, timed at the end of utterances. Additionally, we compute z-scores per subject and feature to account for biological differences in people's voices.

6. Vocalic Model

Here, we elaborate on our instantiation of a coordination model for vocalic and semantic modalities. We model coordination as a continuous variable, with its time evolution described by the following equations:

$$
\begin{aligned}
u_t &= \begin{cases} \mu_{u_0} + \epsilon_{u_0} & t = 0 \\ u_{t-1} + \epsilon_{u_t} & t \neq 0 \end{cases} \\
C_t &= \text{sigmoid}(u_t) \ ,
\end{aligned} \tag{5}
$$

where μ_{u_0} represents the initial coordination value, ϵ_{u_0} and ϵ_{u_t} denote process noise with a prior distribution of $\mathcal{N}(0, \sigma_u^2)$, and u_t serves as an auxiliary variable which is then fed into a sigmoid to constrain coordination (C_t) to lie within $[0, 1]$.

We denote the vocalic component as $n = 1$ and the semantic component as $n = 2$. The latent vocalic component is modeled with four dimensions, each corresponding to a vocalic feature. Further, it is blended using a pairwise blending scheme for serialized data, as outlined in §3.3. Formally, $A_t^{1,p}$ and its observations can be described by the following equations:

$$
\begin{aligned}
A_t^{1,p} &= \begin{cases} \mu_{A_0}^{1,p} + \epsilon_{A_0^{1,p}} & t = 0 \\ \left(1 - C_{f^1(t)}\right) A_{t-1}^{1,p} + C_{f^1(t)} A_{t-1}^{1,p'} + \epsilon_{A_t^{1,p}} & t \neq 0 \end{cases} \\
O_t^{1,p} &= A_t^{1,p} + \epsilon_{O_t^{1,p}} \ ,
\end{aligned} \tag{6}
$$

where $\mu_{A_0}^{1,p}$ denotes the initial latent value for person p, and p' refers to the previous speaker other than p. The function $f^1(\cdot)$ maps the time in the component scale to the time in coordination scale. The terms $\epsilon_{A_0^{1,p}}$ and $\epsilon_{A_t^{1,p}}$ represent the process noise with a prior distribution of $\mathcal{N}(0, \sigma_{A^1}^2)$, shared across subjects. Moreover, $\epsilon_{O_t^{1,p}}$ stands for the observation noise with a prior distribution of $\mathcal{N}(0, \sigma_{O^1}^2)$, shared among subjects and vocalic features.

In experiments with synthetic data, we estimate all parameters. However, the model has too much power to infer σ_u, σ_{A^1}, and σ_{O^1} on real data which is quite noisy and sparse. Thus, we set $\sigma_u = 0.5$ and $\sigma_{O^1} = 0.1$ (values to which the model is not particularly sensitive). We employed a HalfNormal(1) as a hyper-prior for σ_{A^1}, $\mathcal{N}(0, 5)$ for μ_{u_0}, and $\mathcal{N}(0, 1)$ for $\mu_{A_0}^{1,p}$. In experiments where σ_{O^1} and σ_u are not fixed, we use a HalfNormal(1) as their hyper-prior.

Semantic Links We integrate insights from Rational Speech Act models (RSA) (Goodman & Frank, 2016) to introduce an additional modality based on semantic information in team conversations. In RSA, a set of probabilities is often used to model the reasoning processes of speakers and listeners—e.g., the probability of a speaker choosing an utterance.

In our context, each subject acts as both a listener to the previous speaker and a speaker for the subsequent listener at every time step. Semantic link events are identified by analyzing event labels assigned to subjects' utterances, signaling coordination-indicative semantics in temporally proximate utterances from different subjects. In particular, we define a set of pairs of 'source' (s_i) and 'target' (t_i) labels, $\mathcal{L} = \{(s_1, t_1), ..., (s_n, t_n)\}$ that we believe are indicative of team coordination (see Appendix C). We seek pairs of utterances with matching labels, such as (*HelpRequest*, *HelpCommand*), representing a player's request for help and another player's acknowledgment of assistance, respectively.

The time scale of this component is defined by moments when the combination of labels produced by any subject other than p in the past w_u time steps ($w_u = 5$ in our experiments) and the labels produced by p at time t forms any pair in $\mathcal{L}$. Formally, it is the set of times t for which $\exists i \in \{1, \ldots, w_u\}$ such that $\left(\mathcal{U}_{t-i}^{p'} \times \mathcal{U}_t^p\right) \cap \mathcal{L} \neq \emptyset$, where $\mathcal{U}_t^p$ is the set of labels associated with the utterance produced by the subject p ending at time t.

We set the observed values of this component to 1 and model them directly as a function of coordination. In particular, observations are sampled from Gaussian distributions with mean given by the levels of coordination over time. We use the standard deviation of this distribution to control the event's significance: the larger its value, the less likely the event (e.g., response to a help request) is associated with increased coordination as the Gaussian distribution around the coordination level becomes wider. Formally,

$$O_t^{2,p} = C_{f^2(t)} + \epsilon_{O_t^{2,p}}, \tag{7}$$

where $\epsilon_{O_t^{2,p}}$ is the observation noise with a prior distribution $\mathcal{N}(0, \sigma^2_{O^2})$, and the function $f^2(\cdot)$ maps the time in the component scale to the time in the coordination scale. In experiments with real data, we set $\sigma^2_{O^2}$ to 5.0.

The models with semantic links, vocalics and vocalics plus semantic links as modalities are denoted as M_{link}, M_{voc} and $M_{\text{voc+link}}$ respectively. Additionally, we infer coordination at every second during the 17-minute missions, resulting in a time series of $T = 1020$ points. We chose one second as our temporal resolution due to the typical duration of utterances during missions which ranges from a few seconds to a few tens of seconds.

7. Evaluation

Coordination is an intuitive concept, but we are not aware of a general computational definition in the literature. Comparing our model with other methods, invariably designed for particular data, is challenging, as ground truth for coordination is not known or even well-defined. Evaluation is thus not trivial as we are developing a methodology for measuring a concept while simultaneously exploring when the concept is predictive.

In this work, we employ posterior predictive analysis to show that a model incorporating coordination yields improved predictions of future observations. We choose a window size, $w = 5$, and construct a set of 10 time points, τ, comprising random integers uniformly sampled without replacement from the interval $(T/2, T - w)$, where T represents the size of the coordination scale. For each $t \in \tau$, we fit a model up to time t and use 100 samples from the posterior distribution as initial particles for the system. Subsequently, we evolve the system from time $t + 1$ to time $t + w$ (held out) with ancestral sampling and compare the samples $\hat{O}_t$ with real observations O_t, within that interval. This procedure provides a Monte Carlo estimate of the expected squared error of predictions in the future with respect to the model's posterior distribution. We then calculate the root mean squared error (RMSE) and standard error of the mean across different missions in each dataset. Formally,

$$\text{RMSE}(\tau, w) = \sqrt{\frac{1}{wL|\tau|} \sum_{l=1}^{L} \sum_{t \in \tau} \sum_{t'=t+1}^{t+w} \left(\mathbb{E}\left[\hat{O}_{t'}^{1,l}\right] - O_{t'}^{1,l} \right)^2}, \tag{8}$$

where $l \in \{1, 2, 3, 4\}$ indexes the vocalic features and L is the number of vocalic features used (i.e., $L = 4$). For simplicity, we omit the subject's superscript.

We validate this evaluation criteria, explore model properties and inference using synthetic data. Following this, we assess real data, demonstrating that a model with coordination consistently results in smaller RMSEs. Finally, we gauge the predictive power of estimated coordination on team performance across various conditions compared to baseline predictors.

Experiments with Synthetic Data We use a smaller version of M_{voc} with just pitch and intensity to generate synthetic data using ancestral sampling. For data generation, we set $\sigma_{A^1} = 0.01$, $\sigma_{O^1} = 1.0$, and fix the value of coordination to different levels to check that we can recover them using the chosen inference procedure.

In Figure 2, coordination determines the rate at which subjects' pitches converge to a shared value. Note that this represents an extreme scenario with persistent coordination and no noise in the process or observations (plots feature σ_{A^1} and σ_{O^1} set to 0 for clarity). While perfect coordination is unlikely in the real world, this scenario helps elucidate our model's hypotheses in a noise-free context.

Crucially, the figure highlights we can estimate the inherent coordination level employed in generating the noisy data. Data was generated fixing coordination to be constant, but M_{voc} assumes coordination changes over time, causing increased uncertainty the longer in the future we try to estimate coordination. When subjects have reached shared vocalic feature values, any coordination level is acceptable. However, the accurate expected value is maintained due to the early time steps significantly contributing to inferring the rate of convergence in the subjects' vocalic features.

Next, we compute the RMSEs of the predicted observations to assess whether this measure is a suitable evaluation criterion for detecting signs of coordination in the data. Specifically, we establish two models: (i) M_c, with positive coordination, and (ii) M_x, which is an ablated version of M_c without the coordination variable. Subsequently, we generate three datasets D_c, D_x, D_r with 10 samples each (each sample has 50 time steps). D_c and D_x, are generated by drawing samples from M_c and M_x respectively, and D_r by generating iid samples from a standard normal distribution.

Table 1: RMSE and standard error (in parentheses) of predicted vocalics on random data (D_r) and synthetic data generated by models with (M_c) and without coordination (M_x).

	D_x	D_c	D_r
M_x	0.05 (.002)	0.18 (.004)	1.05 (.003)
M_c	0.06 (.003)	0.11 (.004)	1.04 (.003)

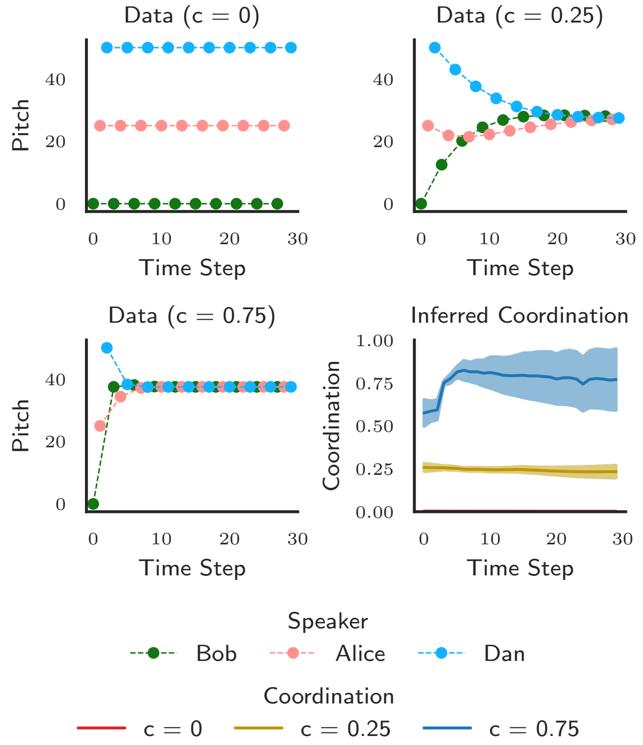

Figure 2: Synthetic data and inferred coordination in M_{voc}. In the absence of coordination, the pitches of Bob, Alice, and Dan evolve independently. As coordination intensifies, their pitches converge, indicating an influence on each other's vocal styles. For clarity, the visualizations display data before introducing noise. The bottom-right plot depicts the inferred coordination for the generated datasets after noise is added.

We then fit both models on the three datasets and compute RMSEs.

Results in Table 1 showcase the superiority of M_c in predicting future data when D_c is generated from a model with coordination set to 0.2. Similar outcomes were observed for various coordination levels. Notably, in the absence of coordination signals in the data, M_c performs similarly to M_x, as it performs no better than chance on D_r and it can effectively learn coordination equal to 0 on D_x. In such cases, modeling coordination may not be advantageous, given the additional computational resources required for inference and the lack of meaningfulness in the estimated value.

Table 2: RMSE and standard error (in parentheses) of predicted vocalics on data from the ASIST and ToMCAT datasets.

	ASIST	ToMCAT
$M_{\text{x,real}}$	1.59 (.05)	1.45 (.04)
M_{voc}	1.45 (.05)	1.32 (.04)
$M_{\text{voc+link}}$	1.36 (.03)	1.34 (.07)

Experiments with Real Data To evaluate the proposed approach on real data, we use M_{voc}, M_{link} and $M_{\text{voc+link}}$ to model teams of three humans collaborating on the complex, time-constrained urban search and rescue (USAR) missions from the ASIST and ToMCAT datasets.

We begin by evaluating RMSEs for M_{voc}, $M_{\text{voc+link}}$, and a modified version of the former that lacks the coordination variable ($M_{\text{x,real}}$). All models share equal initializations for the common variables. The main idea is that if $M_{\text{x,real}}$ outperforms the other models, we cannot use the inferred coordination to draw any conclusions about the teams, as there would be no evidence that the data reflects an underlying coordination process. However, results in Table 2 demonstrate otherwise—both M_{voc} and $M_{\text{voc+link}}$ models outperform $M_{\text{x,real}}$ in predicting the data. Furthermore, the inclusion of the semantic modality leads to improved performance in the ASIST dataset. As semantic links are conditionally dependent solely on coordination (refer to §6), we infer that the enhanced prediction is attributed to a more accurate estimation of the underlying coordination level.

Without assuming a direct association between coordination and final team performance, we explore whether estimated coordination predicts the final team score across different trials and conditions. Here we split the data into training and test splits using leave-one-out cross-validation (LOOCV). We fit linear regressor on the training set, using the average of coordination level peaks (see Appendix D) as input features and the final team score as the target value. We then use the fitted model to predict the final team score in the test split and quantify the performance of the model using the mean absolute error.

We compare the results against two baseline models: (i) M_{avg}, which predicts the score in the test split by averaging scores from the training split, and (ii) M_{hyp}, which fits a linear hyperplane using team scores as target values and 4D-vectors formed by the mean of each vocalic feature across subjects per mission as input features.

We summarize final team score prediction results in Table 3. We find that coordination is predictive of team score in the second mission, but not in the first. Coordination is about working together—not doing so is more likely to result in no correlation than a negative one. We hypothesize that this is due to teams still learning how to play the game in the first mission, whereas their interactions are more strategy-driven in the second mission. Indeed, a pairwise t-test (per team) on the final score in missions A and B shows that the score is larger in the second mission (p = .001 in the ASIST dataset

Table 3: Mean absolute error of final team score predictions on the ASIST and ToMCAT datasets, with standard error of the mean in parentheses. Results were computed using LOOCV.

		ASIST			ToMCAT
Mission	Model	No-Advisor	Advisor	Combined	Advisor
A	M_{avg}	142 (30)	106 (18)	120 (17)	121 (21)
	M_{hyp}	168 (18)	147 (40)	129 (21)	110 (36)
	M_{voc}	178 (45)	90 (21)	120 (19)	139 (24)
	M_{link}	160 (32)	119 (21)	125 (17)	132 (24)
	$M_{\text{voc+link}}$	174 (42)	92 (19)	117 (19)	146 (22)
B	M_{avg}	113 (19)	179 (25)	143 (17)	104 (22)
	M_{hyp}	151 (29)	147 (28)	150 (20)	86 (24)
	M_{voc}	93 (12)	125 (26)	112 (18)	74 (17)
	M_{link}	117 (21)	192 (27)	146 (20)	125 (30)
	$M_{\text{voc+link}}$	102 (13)	101 (20)	105 (14)	106 (18)

and p = .004 in the ToMCAT dataset).

Notably, using just the semantic modality did not lead to better performance (M_{link}), but using both semantic and vocalic data ($M_{\text{voc+link}}$) appears to enhance the predictive accuracy of final team scores in the ASIST dataset under the Advisor condition, but not in the ToMCAT dataset. These findings align with the results observed in Table 2, where the model with semantic information yielded smaller predictive error compared to the vocalic-only model in the ASIST dataset but not in the ToMCAT dataset. Therefore, these results strengthen the validity of using data prediction errors as a metric to ascertain the meaningfulness of coordination estimates.

8. Discussion

We propose a novel computational approach to interpersonal coordination, treating it as a latent variable encoding how the latent states of one person influence another. This is crucial as phenomena in interpersonal interaction manifest across multiple data modalities and time scales, such as alterations in vocal behaviors, body language, neural activity, and semantic reciprocation during collaborative tasks.

Our approach offers a general and intuitive framework for capturing dynamic causal processes, representing coordination as the degree of mixing of evolving distributions. This excludes synchrony that is the result of a common cause from a shared environment and it is less brittle than capturing similar behavior via coupled linear dynamical systems (Guan et al., 2015).

Introducing a new measure for coordination, which consolidates complex phenomena, poses evaluation challenges due to the absence of a ground truth. The ongoing process involves ensuring the measure yields consistent conclusions, is predictive of outcomes, and leads to new theories and experiments. Future work aims to apply our definition to evaluate coordination in diverse settings and modalities, including posture, facial expressions, gaze, core physiological response, and brain scan data. Additionally, we plan to address antisymmetric coordination, recognizing situations where team members deliberately explore different options, providing valuable insights for AI team coaches.

9. Limitations

We are simultaneously defining and evaluating coordination, which makes evaluation methodologically challenging. We have designed the system to infer coordination where there is actual causal influence and not to pick up spurious coordination due to a common cause (e.g., the spatially separated movie goers), however we did not study how robust we are to spurious coordination. We also did not explore non-linear dependencies between coordination and team score, which might be fruitful because the benefits of coordination often are 'U-shaped' (Wiltshire et al., 2019), as collaborators being too similar can be suboptimal. Finally, our strategy to identify semantic links between utterances is simplistic, and may have overlooked deeper semantic associations. We did not evaluate this compared to human coders, which is challenging.

Acknowledgements

This research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-20-1-0002 and Cooperative Agreement Number W911NF-24-2-0034. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding

any copyright notation herein.

Impact Statement

This paper presents work contributing to explanatory modeling of human interaction that can support important future applications such as AI-based team coaching, AI-based tools for diagnosing and improving classroom interactions, evaluation of autistic behavior, and automated analysis of multi-model interaction behavior data in support of social science research.

References

Amon, M. J., Vrzakova, H., and D'Mello, S. Beyond Dyadic Coordination: Multimodal Behavioral Irregularity in Triads Predicts Facets of Collaborative Problem Solving. *Cognitive Science*, 43, 10 2019. doi: 10.1111/cogs.12787.

Boersma, P. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

Borrie, S. A., Lubold, N., and Pon-Barry, H. Disordered speech disrupts conversational entrainment: a study of acoustic-prosodic entrainment and communicative success in populations with communication challenges. *Frontiers in Psychology*, 6, 2015. ISSN 16641078. doi: 10.3389/fpsyg.2015.01187.

Borrie, S. A., Barrett, T. S., Willi, M. M., and Berisha, V. Syncing up for a good conversation: A clinically meaningful methodology for capturing conversational entrainment in the speech domain. *Journal of Speech, Language, and Hearing Research*, 62(2), 2019. ISSN 10924388. doi: 10.1044/2018_JSLHR-S-18-0210.

Brooks, S. and Gelman, A. General Methods for Monitoring Convergence of Iterative Simulations. *J. Comput. Graphi. Stat.*, 7:434–455, 12 1998. doi: 10.1080/10618600.1998.10474787.

Butler, E. A. Coordination in interpersonal systems. *Cognition and Emotion*, 36(8):1467–1478, 2022. doi: 10.1080/02699931.2023.2168624. URL https://doi.org/10.1080/02699931.2023.2168624.

Butner, J., Berg, C., Baucom, B., and Wiebe, D. Modeling Coordination in Multiple Simultaneous Latent Change Scores. *Multivariate Behavioral Research*, 49:554–570, 11 2014. doi: 10.1080/00273171.2014.934321.

Cooper, R. *Coordination Games*. Cambridge University Press, 1999. URL https://EconPapers.repec.org/RePEc:cup:cbooks:9780521570176.

Cornejo, C., Cuadros, Z., Morales, R., and Paredes, J. Interpersonal coordination: Methods, achievements, and challenges. *Frontiers in Psychology*, 8, 9 2017. ISSN 16641078. doi: 10.3389/fpsyg.2017.01685.

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-georges, C., Viaux, S., and Cohen, D. Interpersonal Synchrony: A Survey of Evaluation Methods across Disciplines. *IEEE Transactions on Affective Computing*, 3: 349–365, 7 2012. doi: 10.1109/T-AFFC.2012.12.

Engome Tchupo, D. and Macht, G. A. Entropy for team communication pattern recognition. *Applied Ergonomics*, 111:104038, 2023. ISSN 0003-6870. doi: https://doi.org/10.1016/j.apergo.2023.104038. URL https://www.sciencedirect.com/science/article/pii/S0003687023000765.

Eyben, F., Wöllmer, M., and Schuller, B. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pp. 1459–1462, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi: 10.1145/1873951.1874246. URL https://doi.org/10.1145/1873951.1874246.

Fernández, P., Estrada, E., Ollero, M., and Ferrer, E. Dynamical Properties and Conceptual Interpretation of Latent Change Score Models. *Frontiers in Psychology*, 12: 696419, 7 2021. doi: 10.3389/fpsyg.2021.696419.

Fujiwara, K. and Daibo, I. Evaluating interpersonal synchrony: Wavelet transform toward an unstructured conversation. *Frontiers in psychology*, 7:516–516, 2016. ISSN 1664-1078. doi: 10.3389/fpsyg.2016.00516. This article was submitted to Movement Science and Sport Psychology, a section of the journal Frontiers in Psychology.

Fusaroli, R., Bjørndahl, J. S., Roepstorff, A., and Tylén, K. A heart for interaction: Shared physiological dynamics and behavioral coordination in a collective, creative construction task. *Journal of Experimental Psychology: Human Perception and Performance*, 42, 2016. ISSN 19391277. doi: 10.1037/xhp0000207.

Gelman, A. and Rubin, D. B. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457 – 472, 1992. doi: 10.1214/ss/1177011136. URL https://doi.org/10.1214/ss/1177011136.

Goodman, N. and Frank, M. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20, 09 2016. doi: 10.1016/j.tics.2016.08.005.

Guan, J., Simek, K., Brau, E., Morrison, C. T., Butler, E., and Barnard, K. Moderated and Drifting Linear Dynamical Systems. In Bach, F. R. and Blei, D. M. (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France,* *6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 2473–2482. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/guan15.html.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. Intersubject Synchronization of Cortical Activity During Natural Vision. *Science (New York, N.Y.)*, 303:1634–40, 04 2004. doi: 10.1126/science.1089506.

Henning, R. and Korbelak, K. Social-psychophysiological compliance as a predictor of future team performance. *PSYCHOLOGIA -An International Journal of Psychology in the Orient*, 48:84–92, 6 2005. doi: 10.2117/psysoc.2005.84.

Hoffman, M. D. and Gelman, A. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014. doi: 10.5555/2627435.2638586. URL https://dl.acm.org/doi/10.5555/2627435.2638586.

Huang, L., Freeman, J., Cooke, N., Colonna-Romano, J., Wood, M. D., Buchanan, V., and Caufman, S. J. Artificial Social Intelligence for Successful Teams (ASIST) Study 3, 2022a. URL https://doi.org/10.48349/ASU/QDQ4MH.

Huang, L., Freeman, J., Cooke, N., Colonna-Romano, J., Wood, M. D., Buchanan, V., and Caufman, S. J. Exercises for Artificial Social Intelligence in Minecraft search and rescue for teams, 5 2022b. URL osf.io/jwyvf.

Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A.-L., de Mooij, S. M., Moutoussis, M., Goodyer, I. M., Bullmore, E., Jones, P. B., Fonagy, P., Lindenberger, U., and Dolan, R. J. Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience*, 33:99–117, 2018. ISSN 1878-9293. doi: https://doi.org/10.1016/j.dcn.2017.11.007. URL https://www.sciencedirect.com/science/article/pii/S187892931730021X. Methodological Challenges in Developmental Neuroimaging: Contemporary Approaches and Solutions.

Klopack, E. and Wickrama, K. Modeling Latent Change Score Analysis and Extensions in Mplus: A Practical Guide for Researchers. *Structural Equation Modeling: A Multidisciplinary Journal*, 27:1–14, 4 2019. doi: 10.1080/10705511.2018.1562929.

Knight, A. P., Kennedy, D. M., and McComb, S. A. Using recurrence analysis to examine group dynamics. *Group Dynamics*, 20, 2016. ISSN 10892699. doi: 10.1037/gdn0000046.

Lee, C. C., Katsamanis, A., Black, M. P., Baucom, B. R., Christensen, A., Georgiou, P. G., and Narayanan, S. S. Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions. *Computer Speech and Language*, 28, 2014. ISSN 08852308. doi: 10.1016/j.csl.2012.06.006.

Levitan, R. and Hirschberg, J. Measuring Acoustic-Prosodic Entrainment with Respect to Multiple Levels and Dimensions. pp. 3081–3084, 08 2011. doi: 10.21437/Interspeech.2011-771.

Litman, D., Paletz, S., Rahimi, Z., Allegretti, S., and Rice, C. The teams corpus and entrainment in multi-party spoken dialogues. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2016. doi: 10.18653/v1/d16-1149.

Lubold, N. and Pon-Barry, H. Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. *MLA 2014 - Proceedings of the 2014 ACM Multimodal Learning Analytics Workshop and Grand Challenge, Co-located with ICMI 2014*, pp. 5–12, 11 2014. doi: 10.1145/2666633.2666635. URL http://dx.doi.org/10.1145/2666633.2666635.

Marwan, N., Thiel, M., and Nowaczyk, N. Cross Recurrence Plot Based Synchronization of Time Series. *Nonlinear Processes in Geophysics*, 9, 5 2002. doi: 10.5194/npg-9-325-2002.

Mathieu, J. E., Luciano, M. M., D'Innocenzo, L., Klock, E. A., and LePine, J. A. The Development and Construct Validity of a Team Processes Survey Measure. *Organizational Research Methods*, 23(3), 2020. ISSN 15527425. doi: 10.1177/1094428119840801.

Miao, G. Q., Dale, R., and Galati, A. (mis)align: a simple dynamic framework for modeling interpersonal coordination. *Scientific Reports*, 13, 10 2023. doi: 10.1038/s41598-023-41516-4.

Moulder, R., Duran, N., and D'Mello, S. Assessing Multimodal Dynamics in Multi-Party Collaborative Interactions with Multi-Level Vector Autoregression. pp. 615–625, 11 2022. doi: 10.1145/3536221.3556595.

Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN 0262018020.

Nitschke, R., Wang, Y., Chen, C., Pyarelal, A., and Sharp, R. Rule Based Event Extraction for Artificial Social Intelligence. In *Proceedings of the First Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*, pp. 71–84, Gyeongju, Republic of Korea, 10 2022. International Conference on Computational

Linguistics. URL https://aclanthology.org/2022.pandl-1.9.

Pyarelal, A., Duong, E., Shibu, C. J., Soares, P., Boyd, S., Khosla, P., Pfeifer, V., Zhang, D., Andrews, E. S., Champlin, R., Raymond, V. P., Krishnaswamy, M., Morrison, C., Butler, E., and Barnard, K. The tomcat dataset. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Rahimi, Z., Kumar, A., Litman, D., Paletz, S., and Yu, M. Entrainment in multi-party spoken dialogues at multiple linguistic levels. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2017-August, 2017. doi: 10.21437/Interspeech.2017-1568.

Rahimi, Z., Litman, D., and Paletz, S. Acoustic-prosodic entrainment in multi-party spoken dialogues: Does simple averaging extend existing pair measures properly? In *Lecture Notes in Electrical Engineering*, volume 510, 2019. doi: 10.1007/978-3-319-92108-2_18.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.*, 2:e55, 2016. doi: 10.7717/peerj-cs.55. URL https://doi.org/10.7717/peerj-cs.55.

Schmidt, R. C., Morr, S., Fitzpatrick, P., and Richardson, M. J. Measuring the Dynamics of Interactional Synchrony. *Journal of nonverbal behavior*, 36(4):263–279, 2012. ISSN 0191-5886. doi: 10.1007/s10919-012-0138-5.

Strang, A. J., Funke, G. J., Russell, S. M., Dukes, A. W., and Middendorf, M. S. Physio-behavioral coupling in a cooperative team task: contributors and relations. *Journal of experimental psychology. Human perception and performance*, 40(1):145–158, 2 2014. ISSN 1939-1277. doi: 10.1037/A0033125. URL https://pubmed.ncbi.nlm.nih.gov/23750969/.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

von Holst, E. *The Behavioural Physiology of Animals and Man: The Collected Papers of Erich Von Holst*. Number v. 1 in The Behavioural Physiology of Animals and Man: The Collected Papers of Erich Von Holst. University of Miami Press, 1973. ISBN 9780870242618. URL https://books.google.com/books?id=4FQXAQAAIAAJ.

Wallot, S. Multidimensional Cross-Recurrence Quantification Analysis (MdCRQA) – A Method for Quantifying Correlation between Multivariate Time-Series. *Multivariate Behavioral Research*, 54(2):173–191, 2019. doi: 10.1080/00273171.2018.1512846. URL https://doi.org/10.1080/00273171.2018.1512846.

Walton, A. E., Richardson, M. J., Langland-Hassan, P., and Chemero, A. Improvisation and the self-organization of multiple musical bodies. *Frontiers in psychology*, 6: 313–313, 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.00313. Edited by: Adam M. Croom, University of Pennsylvania, USA.

Washburn, A., DeMarco, M., de Vries, S., Ariyabuddhiphongs, K., Schmidt, R. C., Richardson, M. J., and Riley, M. A. Dancers entrain more effectively than non-dancers to another actor's movements. *Frontiers in Human Neuroscience*, 8:800, 10 2014. ISSN 16625161. doi: 10.3389/fnhum.2014.00800.

Wiltshire, T., van Eijndhoven, K., Hałgas, E., and Gevers, J. Prospects for Augmenting Team Interactions with Real-Time Coordination-Based Measures in Human-Autonomy Teams. *Topics in Cognitive Science*, 3 2022. doi: 10.1111/tops.12606.

Wiltshire, T. J., Butner, J. E., and Fiore, S. M. Problem-Solving Phase Transitions During Team Collaboration. *Cognitive Science*, 42(1), 2018. ISSN 15516709. doi: 10.1111/cogs.12482.

Wiltshire, T. J., Steffensen, S. V., and Fiore, S. M. Multiscale movement coordination dynamics in collaborative team problem solving. *Applied Ergonomics*, 79:143–151, 9 2019. ISSN 0003-6870. doi: 10.1016/J.APERGO.2018.07.007.

Zhang, M., Beetle, C., Kelso, J. A. S., and Tognoli, E. Connecting empirical phenomena and theoretical models of biological coordination across scales. *Journal of the*

Royal Society Interface, 16, 2019. ISSN 17425662. doi: 10.1098/rsif.2019.0360.

A. Inference Details

We implemented the models using PyMC 5.0.2 (Salvatier et al., 2016), with 2000 warm-up iterations, 2000 samples, 4 parallel chains, a target acceptance probability of 0.9 and default values for the remaining parameters. Convergence was assessed by ensuring that $\hat{R} < 1.1$ (Gelman & Rubin, 1992; Brooks & Gelman, 1998) for the latent variables in the model.

All experiments were run on a machine with 128 AMD EPYC 7542 CPU cores and inference runs took ≈ 10 minutes per trial on average.

B. Virtual Search and Rescue

We provide a brief description of the task and salient features of the ASIST and ToMCAT datasets. For more details, see the study preregistration (Huang et al., 2022b) and the documentation for the datasets (Huang et al., 2022a; Pyarelal et al., 2023).

B.1. Task

Both the ASIST and the ToMCAT datasets were collected from participants playing the same task. The task consists of a collapsed building where a team of three participants have to collaborate in different ways to find, treat and move victims to safe areas in order to earn score points. The scenario is implemented in Minecraft, and the testbed is publicly available at `https://gitlab.com/artificialsocialintelligence/study3`.

Each participant is assigned a different role that entails different abilities and speeds. All participants can carry victims, but the *medic* is the only one that can treat them. The *engineer* is the only one that can remove obstacles. The *transporter* has the highest speed; hence it can explore the areas faster. In addition, all participants have a set of markers (SOS, rubble, threat, regular victim, critical victim, no victim, type B, type A) that they can use unlimited times to communicate their discoveries with the other players.

Regular victims (worth 10 points) and *critical* victims (worth 50 points) are scattered around the building under piles of rubble or inside rooms. A regular victim can be injured with abrasions (type A) or bone damage (type B), and a critical victim (type C) can only be treated after it is stabilized, which happens when more than one participant arrive at its vicinity. To earn the points associated with treated victims, participants must move them to safe areas matching their types. The victim types are only known by the medic. Therefore, communication and coordination are essential for teams to succeed in the mission.

Each team plays two 17-minute-long games: mission A and B, with the same scenario but different victim placement configurations. The first two minutes are reserved for planning: participants can move and talk to each other, but access to the building is blocked. In addition, before the first game, participants play a tutorial mission with small tasks devoted to familiarizing them with their role's specific abilities and game dynamics. After completing all the individual tasks, the team will convene and gain access to a designated portion of the building. This setup enables them to undergo training in a context more closely aligned with the scenarios they will encounter during their upcoming missions. The victim locations during this part of the training phase differ from those in missions A and B.

B.2. Experimental Design

ASIST In the ASIST experiments, participants carried out the Minecraft missions from their respective homes, operating remotely. They were instructed to use headsets, and audio streaming quality was qualitatively checked before proceeding with the mission. Individual audio streams were recorded separately for each team member.

ToMCAT In the ToMCAT experiments, participants were physically situated in a lab and equipped with an array of sensors, encompassing fundamental physiological measurements like EKG (electrocardiogram), skin conductance (GSR), as well as eye trackers, in addition to a comprehensive setup featuring both fNIRS (functional near-infrared spectroscopy) and EEG (electroencephalography) caps. In addition, directional microphones were positioned in front of each participant on the stations they were operating, resulting, also, in the creation of separate audio files for each team member.

Further, prior to playing the Minecraft missions, participants in the ToMCAT experiments had to complete various small tasks. This included a resting phase, as well as participation in individual and team affective tasks, rhythmic finger tapping exercises, and competitive and cooperative virtual ping-pong games. Consequently, the ToMCAT experiments were longer and demanded more diverse cognitive load from the participants compared to the ASIST experiments.

B.3. Data Pre-processing

ASIST After removing trials with audio problems, we were left with data from 30 teams of three participants: 16 teams in the 'No-Advisor' condition and 14 in the 'Advisor' condition. The published data from this experiment included speech transcripts automatically generated by the Google Cloud Speech automatic speech recognition (ASR) service, event labels (Nitschke et al., 2022), raw audio files, as well as a set of vocalic features computed using openSMILE (Eyben et al., 2010).

We noticed that the timestamps in the original ASR transcriptions were not aligned with the utterances in the audio files and their vocalic features. We corrected these misalignments by reprocessing the audio files through: (i) Whisper (Radford et al., 2023) to regenerate transcriptions along with start/end timestamps, and (ii) openSMILE (Eyben et al., 2010) to recompute the vocalic features and associated timestamps. We used the same openSMILE configurations as the ones used to produce the original dataset. We used Whisper instead of the original Google solution because we noticed by manual inspection that the former yielded better transcripts.

In addition, we sent the transcriptions through the event extraction system (Nitschke et al., 2022) to regenerate event labels for the utterances. The transcriptions, event labels, and vocalics in the original data were computed in real-time from live audio streams instead of subsequently saved audio files, which are what we used. Therefore, the new data is mostly like the original one, except for the timestamps and a few values computed in periods of small glitches in the audio files. This reprocessing procedure ensures that the audio files are aligned correctly with detected utterances, associated vocalics, and labels, providing more reliable data for further analysis.

Among the available features produced by openSMILE, we selected `f0final_sma` for pitch, `rmsenergy_sma` for intensity, `jitterlocal_sma` for jitter, and `shimmerlocal_sma` for shimmer.

We manually inspected each identified utterance and associated audio to detect speech quality. During this process, we identified and subsequently removed some utterances with sound contamination from multiple participants, which can cause duplicated utterances in the data and unreliable vocalic features. Specifically, in the 'No-Advisor' condition we removed 53 out of 5339 identified utterances, while in the 'Advisor' condition, we excluded 14 out of 5001 utterances. This contamination can be attributed to the remote nature of the data collection—in some instances, the participants did not comply with the instructions to wear headphones during the experiment, and thus other player's voices were played over their computer speakers and captured by their microphones.

In addition, we observed that the ASR service used occasionally split a single utterance into multiple subsequent ones. To address this issue, we manually merged consecutive utterances from a participant, along with their associated vocalic features and event labels.

ToMCAT From this dataset, we only used data from the 'Advisor' condition, primarily due to the limited number of trials available in the 'No-Advisor' condition. We discarded data from trials with problems in the audio and trials were one of the experimenters had to fill in for one of the participants, to avoid using potentially biased data. In total, we utilized data from 9 different teams.

Similar to the ASIST dataset, one of the products of this dataset is speech transcripts that were automatically generated by the same Google ASR service. However, these transcripts were substantially marred by duplications caused by cross-talk, given the close proximity of participants playing in the same room. Consequently, we resorted to a manual approach for finding utterances within each audio file attributed to the primary speaker. We used Praat (Boersma, 2001) to perform this annotation task.

When defining the boundaries of an utterance, we considered pauses occurring within an utterance as non-silence unless they extended beyond one second, except in cases where it was evident from the audio that the pause signaled a response to a question posed by another participant. In such instances, we divided the utterance into two segments. Ultimately, we generated transcripts using the Whisper ASR system (Radford et al., 2023) and subsequently undertook manual corrections to rectify transcription errors. The generation of vocalics and event extraction followed the same methodology employed during the preprocessing of the ASIST dataset.

The resulting reprocessed data for both datasets with all the different types of data streams is too large to submit as part of the supplementary material. However, the subset of data needed to train and evaluate our models is further derived and much smaller, so we include it along with the code in the interest of reproducibility.

C. Semantic Link Labels

The event extraction system (Nitschke et al., 2022) extracts event labels associated with each detected utterance. We use a subset of these labels to look for temporally separated pairs of utterances that are indicative of team coordination. In particular, we selected six pairs of ('source', 'target') labels to detect semantic linkage between two utterances from different subjects. The specific labels and associated semantics we wanted to capture are shown in Table 4.

Table 4: Pairs of source-target labels used for semantic linking.

Source	Target	Description
HelpRequest	RescueInteractions	Players say they are coming in response to another player's mention of a victim.
HelpRequest	HelpCommand	Players offer assistance in response to another player's request for help.
HelpRequest	Move	Players say they are coming in response to another player's request for help.
HelpRequest	Agreement	Players demonstrate awareness about another player's request for help.
Plan	Agreement	Players demonstrate awareness about another player's plan.
Question	Agreement	Players demonstrate agreement in response to another player's question.

Table 5: Example utterances from the ASIST Study 3 dataset and their extracted event labels indicating semantic link events.

Utterance 1	Utterance 2	(Source, Target) label pair
Medic: *Okay engineer I need your help in K4 they're like to victim here but there's Rubble okay?*	**Engineer**: *All right I'm on the way.*	(HelpRequest, Agreement)
Transporter: *Engineer can you come to D3?*	**Engineer**: *Yeah.*	(Question, Agreement)
Transporter: *C4 all right this is transporter we had a red one in A4 that was marked as an issue but there's nothing there also to engineer I am stuck in D4.*	**Engineer**: *This is engineer and I'm on my way.*	(HelpRequest, Move)

While the event extraction system implemented by Nitschke et al. (2022) is promising, it is not without its limitations. They report achieving an F1 score of 0.94 for simple events and 0.77 for complex events (the evaluation was done on the ASIST dataset). Consequently, it is anticipated that our matching procedure may yield some false positives, and we may have missed some existing semantic link events in the data. Nevertheless, we present in Table 5 a few utterances from the data that illustrate the system's ability to capture meaningful semantics from the experiments employed in our evaluation.

On average, we found 31 ± 22 links per mission in the ASIST dataset, and 39 ± 14 in the ToMCAT dataset.

D. Peaks in Coordination

To identify local maxima in the coordination series, we use the function `signal.find_peaks` from the SciPy library (Virtanen et al., 2020) with the parameter *width* set to 5 to ignore sharp peaks of short duration.

We select the mean coordination at peaks instead of the mean over the whole series for two reasons: (i) coordination in the beginning and at the end of a mission are unlikely to be as representative as levels of coordination in the middle region, where peaks are more prolific, (ii) vocalic features and semantic linkage are sparsely observed. Per mission, on average, we only observe vocalic features on 14% of the time steps in the ASIST dataset and 21% in the ToMCAT dataset; and semantic links on 3% of the time steps in the ASIST dataset and 4% in the ToMCAT dataset. While our approach can naturally handle missing data, the levels of coordination inferred at periods with no evidence are less accurate.