# I Don't Always Do What I Am Told: The Case for Task Incongruence-Aware Systems for Technology-Supported Reflection on Social Interactions

Chenxu Hao
Delft University of Technology
Delft, Netherlands
C.Hao-1@tudelft.nl

Tiffany Matej Hrkalovic
Jheronimus Academy of Data Science
's-Hertogenbosch, Netherlands
t.matej.hrkalovic@tue.nl

Bernd Dudzik
Intelligent Systems/Pattern Recognition & Bioinformatics
Delft University of Technology
Delft, Netherlands
b.j.w.dudzik@tudelft.nl

Hayley Hung
Intelligent Systems
Delft University of Technology
Delft, Netherlands
h.hung@tudelft.nl

## Abstract

Technology-supported Self-Reflection (TSR) systems aim to facilitate learning through self-reflection. In the social domain, these systems often nudge individuals to perform better towards a certain outcome in a targeted task. Typically, this idea of support assumes that users pursue specific goals, which system creators anticipate will govern users' behavior in targeted tasks. However, how goals shape behaviors in social interactions is complex: individuals' goals can be highly idiosyncratic and multifaceted, sometimes contradictory to their best interest, and they are often not explicitly formulated in advance. We believe that viewing this complex landscape through the lens of a simple task formulation imposes limitations on the type of support these systems can provide. In this paper, we take a first step towards investigating the incongruence between assumed task goals and observed users' behaviors and decisions in a partner selection setting, where participants need to select a partner for a cooperation task, and the apparent task goal can co-exist with other personal or relationship-oriented goals. First, we identify task-incongruent decisions, showing that even in a highly structured and incentivized setting, some choices deviate from the assumed task goal. Second, using a Large Language Model (LLM) to analyze conversational behavior, we explore topic-based and interactional indicators of task incongruence. Although the behavioral evidence is mixed, these findings still illustrate that what appears to be a reasonably assumed goal by the system may not be what users end up pursuing. In designing TSR systems, we believe that it is crucial to have capabilities for Task-Incongruence Awareness (TIA), enabling meaningful reflections on how users' consciously or unconsciously self-ascribed goals are aligned with their behaviors and decisions.

This work is licensed under a Creative Commons Attribution 4.0 International License.
*ICMI '26, Napoli, Italy*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2318-6/26/10
https://doi.org/10.1145/3776574.3831154

## CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *Collaborative and social computing theory, concepts and paradigms*; • **Computing methodologies** → *Discourse, dialogue and pragmatics.*

## Keywords

Task Incongruence; Task-Incongruence Awareness; Technology-Supported Reflection; Multimodal User Modeling; Social Interaction; Partner Selection; Large Language Models

**ACM Reference Format:**
Chenxu Hao, Tiffany Matej Hrkalovic, Bernd Dudzik, and Hayley Hung. 2026. I Don't Always Do What I Am Told: The Case for Task Incongruence-Aware Systems for Technology-Supported Reflection on Social Interactions. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26), October 05–09, 2026, Napoli, Italy*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3776574.3831154

## 1 Introduction

*Technology-Supported Reflection (TSR)* systems aim to facilitate users to improve through self-reflection, enabling them to gain new insights and perspectives or change future behaviors by revisiting and rethinking the past [30]. Many TSR systems have been developed to facilitate and augment social interactions among different individuals [10, 20], covering various domains such as communication (e.g., Dressler et al. [14], Kim et al. [26]), social emotional learning (e.g., Afxentiou et al. [1], Akbas et al. [3]), and public speaking (e.g., Goto et al. [18], Huang et al. [23]).

Currently, technologies that support self-reflection for social interaction in these application domains often target specific tasks with clear, pre-defined goals, i.e., they assume certain outcomes are to be achieved through a series of actions [6]. These task-specific systems tend to model the users based on a normative assumption that users will focus on the given task and aim to improve towards these pre-defined goals. Naturally, these systems also aim to help users reflect on their behaviors and decisions with regard to reaching these specific goals [20].

However, in real-life social interactions, the role that goals play in shaping individuals' behavior and self-reflection is often more

complex, multifaceted, and dynamic than assumed by the idea of task formulation. For example,

- *More or Alternate Goals*: people may have other goals than the one(s) assumed by the task formulation. For example, simultaneously pursuing informational and relational motives in conversation [36];
- *Goal Impermanence*: goals may manifest, change, or be abandoned spontaneously during a social interaction (e.g., [31]);
- *Unconventional Behavior*: even when people have the same goal as that of the assumed task, they use completely unexpected strategies to achieve it (e.g., [17]).

As a result, behaviors and decisions in social settings can appear inconsistent with respect to any pre-defined task goals. We name this apparent inconsistency *Task Incongruence*. We posit that designing systems that optimize towards the purported task goal may miss important factors that determine task success, particularly in group-based interactions, or long-term performance [12, 33, 35]. For example, in team collaborations, it is often challenging to disentangle objective task performance and social components of cooperative and collaborative interactions [19].

Subsequently, when TSR systems detect and suggest that people's behaviors do not align with the pre-defined task goal without further awareness, people may reflect and adjust their behaviors and decisions away from goals that could be jointly optimal for the task as well as themselves personally. For example, to maintain high task performance over the long term, people may need to have personal goals that take care of their well-being.

For this reason, we argue in this paper that TSR systems for social interactions should be designed with the capability for *Task-Incongruence Awareness (TIA)*; the ability to detect and intelligently adapt to task incongruent behavior automatically. We believe that enabling applications to adapt to such detections of Task Incongruence will allow them to go beyond nudging people's behaviors and decisions towards a presumed intended outcome in a specific task. Instead, TSR systems could provide feedback for reflection that helps individuals gain deeper insights into how to jointly optimize immediate tasks and personal goals, and to better understand how contextual factors contribute to nuances in successful task completion.

As a first step towards exploring the potential of TIA in providing richer reflection processes on social interactions, we believe it is important to acknowledge and model the complexity of goal-related user behaviors, even in cases where users' motivations seem straightforward. We contribute a novel empirical exploration of the occurrence of task-incongruent behaviors and decisions *even* in a social interaction setting where individuals are explicitly provided with and incentivized towards a defined goal. Specifically, we consider dyadic interactions between strangers for partner selection as captured by the PARSEL [21] dataset. In it, participants were tasked with selecting a partner for a cooperation task where the partner's competence was important, after having a conversation with potential candidates. We argue that this setting is an extreme case of external task influence on the goals present in a conversational interaction. Despite this, we hypothesize that participants' decisions and behavior do not always align with the assumed task goal, illustrating instances of Task Incongruence. To this end, we present three principal contributions in this paper:

(1) A computational exploration of Task Incongruent Social Decisions, where we investigate whether participants make decisions congruent with their perceptions and their alignment with an externally provided task (Investigation 1).
(2) An empirical investigation of Task Incongruence in Social Behaviors, where we explore conversational topics and related interaction patterns as potential behavioral indicators of task-congruency through an LLM-driven analysis (Investigation 2).
(3) A discussion of the implications of these findings for the design of task-incongruence-aware user models and TSR systems in social interaction settings.

## 2 Related Work

We situate our work in the intersection of *TSR system design* for goal setting, and understanding the complexity of goals in social interactions for intelligent systems.

### 2.1 Goal-Setting in TSR Systems

Goal-setting is an important and challenging component for TSR system design. Designers need to determine the intended level of reflection [16] that the system aims to support users to engage in. For example, the level of reflective description aims at understanding and explaining behaviors, whereas the level of transformative reflection aims at perspective changes and intended behavior changes [16]. In addition, while the TSR system may not be designed to support users along all phases of reflection through preparation to action, determining the goal as a preparation for reflection affects how the support system will engage with users [8, 30]. For example, goals of reflection help determine how TSR systems based on personal informatics will collect and integrate data to support users' reflections.

Existing TSR systems in the social interaction domain tend to be task-specific [20]. Given a well-defined targeted task, these systems are commonly designed to promote perspective change or behavior change towards an assumed goal. For example, a TSR system for public speaking may aim to help the users improve a specific set of presentation skills through reflection (e.g., [18, 28, 37]).

Consequently, the evaluation of TSR systems can also be affected by goal-setting. Existing TSR systems are often evaluated based on the quality of the system [10]. Similarly, in the social domain, such support systems are commonly evaluated with user studies focusing on subjective user experience. This reflects a crucial challenge that evaluations of self-reflections in the social domain are not merely about improving objective task performance, but also about gaining additional insights and understanding about users' goals[20].

While TSR systems based on personal informatics often facilitate goal-setting to some extent [10], a recent review shows that very few support systems for self-reflections focus on gaining personal insights or goal-setting in social interactions [20], where people's behaviors are often guided by complex goals. In addition, while some advanced systems approach this complexity through multi-goal or multi-objective optimization [24]–which assumes the

system is aware of and actively optimizing for multiple user goals–this is rarely feasible in social domains. In most cases, the system does not have, and will likely not have, full access to a user's latent internal goals [13].

This suggests that systems designed for self-reflections for social interactions are still limited in how they can navigate the potential existence of multifaceted goals of users. We argue that this lack of complex goal awareness potentially limits the effectiveness of user support in TSR systems.

## 2.2 Task-Incongruent Behaviors and Decisions in Social Interaction

Human social behaviors and decisions are complex, as they are often affected by multifaceted goals that are not merely oriented toward objective tasks. These goals include building social connections and achieving mutual understanding with others [34, 36]. In addition, the complex context of dynamic social interactions also affects social behaviors and decisions. For example, conversations carry not only informational goals, but also relationship-oriented goals [15, 34, 36]. While individuals ideally should translate their goals into their conversation behaviors, they may still make suboptimal choices that appear incongruent to their goals at moments in conversations due to contextual factors such as time pressure and attention [36]. Traditionally, such instances are viewed through the theoretical lens of bounded rationality, which addresses how users' decisions and behaviors can be suboptimal (i.e., bounded) given an explicit goal due to cognitive or informational limits [32]. In this work, by focusing on Task Incongruence, we do not emphasize these suboptimal outcomes or cognitive limits. Rather, we present these incongruent decisions and behaviors–as well as their outcomes–as a reflection point indicating that a user might be intentionally (or implicitly) pursuing a completely different goal that the system is unaware of. Another example is that in team meetings, people often carry different goals. Without a clear mutual understanding of meeting goals, exchanges during meetings can often appear inefficient or task-incongruent [31]. This often requires external support systems to encourage reflections on individual goals, and clarifications and tracking of overarching goals [31].

This indicates that in social settings, objective outcomes of tasks cannot be sufficiently disentangled from social outcomes. Past work has shown that team leaders not only need to be task-oriented to achieve higher group efficacy, but also need to be relationship-oriented to maintain effective communications and foster group cohesion [33], as cohesion is important for the performance and well-being of team members, as well as the effective collaborations in teams [9]. As most TSR systems in the social domain provide support based on the underlying assumption that users ought to change behaviors and decisions for the assumed goal of the targeted task, these systems can become restricted in how they help users to gain insights into their goals and behaviors.

While it is difficult to infer individuals during dynamic settings such as social interactions [6], we believe that designing systems with goal-aware user models that capture users' diverse goals beyond the prescribed task goal is a feasible intermediate step. As a first step towards developing such user models for TSR systems, we explore social decisions and behaviors in a partner selection setting where participants receive clear task instructions and are motivated financially to achieve a prescribed task goal, and we ask if we still observe task-incongruent decisions and behaviors.

# 3 Task Incongruence: Empirical Investigations

## 3.1 Overview

In this section, we describe two empirical investigations of the phenomenon of Task Incongruence in a conversational setting where participating individuals are provided with a clear objective and are strongly incentivized to pursue it. The primary goal of this exploration is to showcase the occurrence (and diversity) of task-incongruent decisions and behaviors in a social setting that is relevant for Technology-Supported Reflection (TSR). For *Investigation 1*, we focus on identifying potential instances of Task Incongruence in *Social Decisions*. These are the choices individuals make based on social interactions they participated in related to their assigned task. In the PARSEL dataset [21] we use, this is the decision of selecting a relevant partner. Building on this, *Investigation 2* explores the potential existence of Task Incongruence in observed *Conversational Behaviors*. In PARSEL, this is operationalized by the range of topics displayed during conversations between zero-acquaintance dyads before they decided on potential collaboration partners. We provide detailed descriptions of our methodology and results for each of these investigations below.

## 3.2 Task Setting and Dataset

The PARSEL dataset is a comprehensive corpus of dyadic social interactions (online) for partner selections and collaborations. The dataset contains 297 participants and measurements of partner selection decisions over several different stages (impression, conversation, cooperation, and collaboration) and includes audiovisual recordings during the interactions, self-reported personality traits, perceptions of others and situations, etc. [21].

In this work, we focus on partner selection for the Joint Competence Task during the impression and conversation stages. The Joint Competence Task is a non-verbal intelligence-based task, requiring participants to solve intelligence-based problems together correctly to gain a monetary reward. This task does not manipulate a conflict of interest among participants but instead assumes a concrete goal in their brief social interactions: to select a competent partner.

In our analysis, we used the 129 participants in PARSEL who were assigned to the Joint Competence Task and whose conversation scripts exist. Every participant in a group of 4-6 participants engaged with others in a round-robin manner through the following processes: 1) *Impression stage*, where each participant reported their subjective perception of others through a first impression based on photographs and made decisions on whether they wanted to select others for the joint task based on selection strength; 2) *Conversation stage*, where each participant had a 3-minute conversation with all other participants in their group, subsequently reported their subjective perception of others, whether they wanted to select their conversation partner for the joint task, and how strongly they wanted to work with that person. In both stages, participants also indicated how competent they perceived their conversation partner to be. Perceived competence was measured by the *mean* rating of how competent, intelligent, and skillful the other person appeared

to be on a scale of 1 to 7, where 1 was *extremely not*, and 7 was *extremely*).

Given these measures, in our experiment, a task-congruent behavior would be to indicate the strongest preference for the person who is perceived to be most competent.

## 3.3 Investigation 1: Exploring Task-Incongruent Social Decisions

Here, we investigate Task Incongruence in the decisions that participants in the PARSEL dataset make when selecting partners for collaboration.

*3.3.1 Measures and Method.* We use the following measures from the PARSEL dataset to analyze Task-Incongruence in Social Decisions: the Perceived Competence scores from PARSEL's Impression stage, and the Perceived Competence scores from the Conversation stage. To ensure we are measuring the Task Incongruence exclusively due to the social interaction, we operationalize the selection strength of perceived competence as the change in selection strength between the conversation and impression stage to discount well-known first impression effects.

We first fit a linear regression model of Change in Selection Strength on Change in Perceived Competence. If a participant does not select another as a partner, then we code the selection strength as 0. Given the results of the regression analysis, we separate the decisions into two categories: task-congruent decisions and task-incongruent decisions. We used studentized residuals to identify instances where changes in partner preference given changes in perceived competence deviated substantially from task-congruent predictions. A threshold of 2 was chosen following standard practice in residual outlier detection. If a point's studentized residual value exceeds this threshold, it was classified as incongruent, while points below the threshold were classified as congruent. All analyses were conducted in Python (3.11.4) and R (4.4.2).

*3.3.2 Analysis and Results.* As shown in Figure 1, Overall, participants' (N=129) perceived competence rating for their conversation partners increased (M=0.55, SD=0.97), and participants' selection strength also increased (M=6.95, SD=39.09) A linear regression indicated that greater increases in perceived competence significantly predicted greater increases in selection strength, $\beta = 15.68, SE = 1.85, t(406) = 8.47, p < 0.001$.

We computed studentized residuals for this linear regression predicting the change in selection strength using the change in perceived competence. Observations with |studentized residual| > 2 were considered as incongruent, as these decisions of partner selection deviated substantially from the predicted pattern. Out of 408 observations from 129 participants, 25 (6%) were classified as incongruent (highlighted in red in Figure 1), while 383 (94%) were classified as congruent.

This shows that the designed task goal was successful; participants tended to choose partners they perceived as more competent. However, some participants exhibited incongruent decisions.

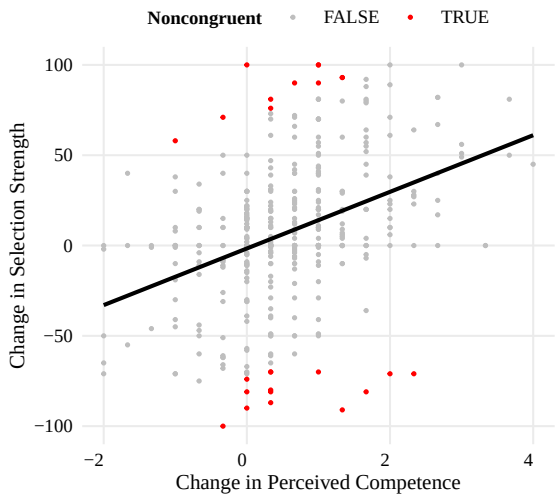



**Figure 1: Change in perceived competence of conversation partner predicts change in selection strength for the Joint Competence Task. Out of 408 observations, 25 instances were classified as task-incongruent.**

## 3.4 Investigation 2: Exploring Task-Incongruent Social Behavior

Our previous exploration points out that instances of Task Incongruence in social interactions can occur even in the strongly constrained circumstances of an experiment. Building on this, we describe a follow-up investigation that specifically explores the relationship between Task Incongruence and the presence of specific Conversational Behaviors in the PARSEL dataset. Our aim here is to better understand how task incongruence manifests itself in behavioral indicators that could be leveraged for relevant multimodal user modeling by intelligent systems.

Prior work suggests that conversation goals can be characterized along two key dimensions: 1) information exchange and 2) relationship building, and individuals may prioritize these dimensions differently depending on the context [36]. For instance, seeking advice can be both highly informational and highly relational. In the context of PARSEL, the assumed task goal is for participants to use the conversation as an opportunity to exchange relevant information and select a competent partner [21]. Thus, we assume that a conversation aligned with this goal would likely focus on task-relevant topics such as their skills, understanding of the Joint Competence Task, or problem-solving approaches. In contrast, Task Incongruence may manifest in Conversation Topics during these interactions that reflect goals other than evaluating the conversation partner's competence, such as establishing a personal connection.

*3.4.1 Measures and Method.* To verify whether Task Incongruence in Conversation Topics is observable, we used automated transcriptions (generated using Whisper X [5]) from the *Conversation Stage*, segmented into speaker turns from the PARSEL dataset. After excluding 2 incomplete transcriptions, we conducted analyses on all remaining available data (217 conversations, corresponding to 382 partner selection observations).

**Table 1: Definitions of conversation topic categories used in the annotation scheme.**

| Topic Category | Definition |
|---|---|
| *Crowdwork Context* | Talk about payment, surveys, or the broader crowdwork platform context (e.g., Prolific, MTurk). |
| *Experimental Task* | Talk directly about the specific puzzle or task in the experiment. |
| *Geographic Locations* | Talk about places people come from, live in. |
| *Hobbies & Leisure* | Talk about hobbies, interests, etc. |
| *Personal Life* | Talk about personal background, daily life, family, routines, food, health, or other personal experiences not primarily framed as work or study. |
| *Socializing (Greetings)* | Greetings, ice-breakers, pleasantries... Short relational exchanges such as asking how someone is doing. |
| *Technical Platform Issues* | Talk about technical problems such as software, hardware, connection, audio/video, or interface issues. |
| *Weather* | Talk explicitly about weather conditions or climate-related small talk. |
| *Work & Education* | Talk about jobs, career background, school, university, training, or academic experience. |

*Coding Scheme Development.* We devised categories of Conversation Topics with a data-driven method on the extracted transcripts. We utilized a local LLM (Ollama [29] running Llama 3 8B [2]) via the LangChain Python library [11] to perform an initial open-ended exploration of a random subset of 20 conversation transcriptions. In this exploration phase, the LLM was given 20 conversation one by one, segmented by turns (start and end time) and speakers. For each conversation, the LLM was prompted to analyze the conversation and segment it by topic changes. The prompt (see Appendix A.1) specified that for each topic, the LLM should provide a high-level category that represented broad themes such as "Hobby" and "Location" and a fine-grained sub-topic that had more specific details. In this work, we focus on the high-level category only.

Based on these preliminary outputs that were also manually reviewed, we refined a list of 9 distinct allowed topics for annotating the topics of all conversations. The categories and their definitions are shown in Table 1.

*Automated Annotation Pipeline.* After the initial exploration of topics based on the subset of conversations, we used the same local LLM (Ollama [29] running Llama 3 8B [2]) via the LangChain Python library [11] to annotate the topics for the entire 217 conversations. To preserve conversational context, the LLM processed the conversations one by one, segmented by turns and speakers. For each conversation, the LLM was instructed to classify every topic into one of the categories from the list of allowed topics, marking the turn and the speaker for when the topic started. To avoid confusion, the LLM was given specific definitions for the topic categories *Experimental Task*, *Technical Platform Issues*, and *Crowdwork Context*. The complete prompt is provided in Appendix A.2. In addition, if a topic did not fit into any of the topics from the list, the LLM was asked to mark the category as *Other* and provide a descriptive subtopic. For each conversation, the LLM output included the following variables: conversation ID, speaker, topic category, start turn, and a descriptive subtopic, in case the topic was classified as *Other*. We focus only on the list of topics in the following analyses.

*Post-Processing.* Following the turn-level annotation, we aggregated contiguous turns of the same topic category into coherent topic episodes. For each episode, we recorded the start turn, end turn, and calculated the total duration for each speaker by summing the timestamps of the included turns corresponding to that speaker. This allowed us to analyze the duration of specific conversational topics in each conversation.

*Validation & Reliability.* Two human annotators (a subset of the authors) coded a validation set independently of one another. It comprised a random sample of 6 conversations, containing a total of 294 turns. For each turn, the human annotators were instructed to assign a topic from the allowed list of topics based on the conversation text. If the text of that turn did not fit into any of the topics from the list, the annotators were asked to mark the topic as *Other*. The inter-rater reliability between the human annotators was substantial ($\kappa = 0.79, p < .001$), indicating a good level of consistency.

To validate the automated annotation, we compared the LLM's topic annotations against the human consensus labels (242 turns) on the validation set (294 turns). The inter-rater reliability between human consensus labels and the LLM annotations was moderate ($\kappa = 0.43, p < 0.001$). In addition, we tested the performance of the LLM model, which achieved a Macro F1-score of 0.58, with a Macro Precision of 0.56 and a Macro Recall of 0.58, indicating moderate topic classification performance across categories.

A detailed confusion matrix showing topic-wise agreement is provided in Appendix B. Inspection of errors revealed that disagreements were primarily concentrated among topics *Experimental Task*, *Geographic Location*, and *Crowdwork Context*. These results reflected that the model captured the core semantic structures of the conversation, while it sometimes diverged on potentially ambiguous cases (e.g., *Geographic Location*) or semantically similar topics (e.g., confusing *Experimental Task* with *Crowdwork Context*).

*3.4.2 Analysis and Results.* We first analyzed the composition of automatically extracted conversation topics per conversation. These topics reflected a mixture of themes that correspond to both task-congruent goals and other goals, such as relationship-building. Specifically, as shown in Figure 2, talks about *Experimental Task* accounted for 9.04% of all turns on average in a conversation, suggesting that participants focused not just on the task objective. Relationship-oriented topics were more prevalent, with *Socializing Greetings* and *Personal Life* constituting 40.93% of the turns, suggesting that participants actively sought social connections despite the designed task goal being to find a competent partner who would do well on the non-verbal intelligence-based task.

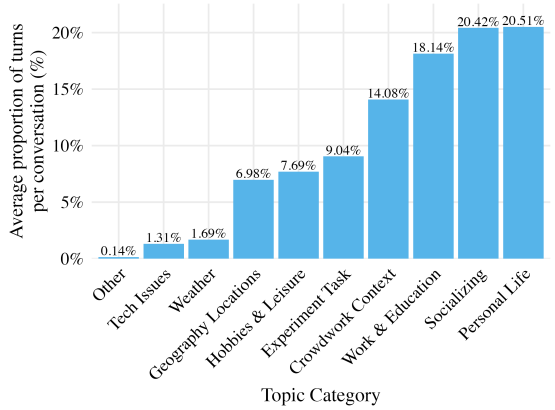



**Figure 2: Proportion of turns on each topic (per interaction).**

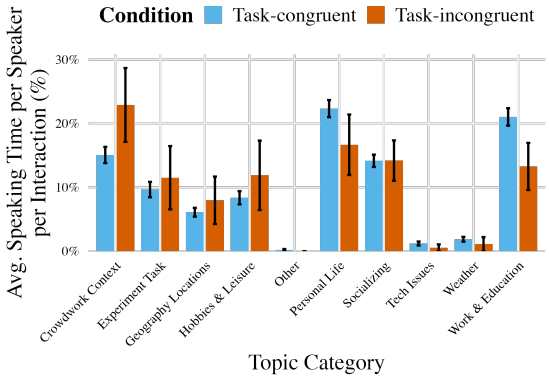



**Figure 3: Average time spent on each topic (per speaker in an interaction), by Task-congruent/incongruent observations.**

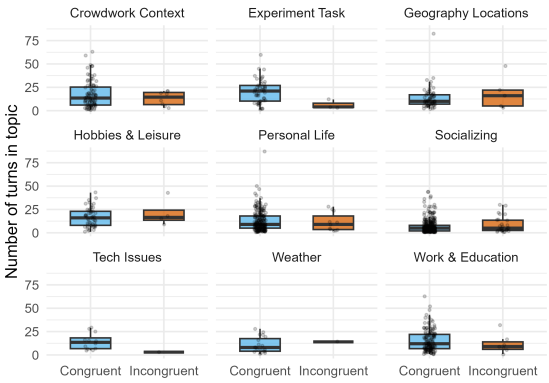



**Figure 4: Exploratory analysis of topic strength across topic categories. Panels show the distribution of topic strength, operationalized as the number of turns a topic lasted after being initiated, separated by congruency.**

We examined the average speaking time people spent on each topic in interactions that resulted in task-incongruent and task-congruent decisions (Figure 3). To test whether the proportion of speaking time allocated to different topics varied by congruency (as referred to Section 3.3), we fitted an ordered beta mixed-effects model in R [7] using glmmTMB, with topic category, congruency, and their interaction as fixed effects, and random intercepts for conversation and speaker-within-conversation. This model was chosen as the dependent variable was a bounded proportion and contained exact zeros.

The model revealed a significant main effect of topic category, $\chi^2(8) = 418.97$, $p < .001$, indicating that speakers distributed their speaking time unevenly across topics. In contrast, there was no significant main effect of congruency, $\chi^2(1) = 1.90$, $p = .168$, and no significant interaction between topic category and congruency, $\chi^2(8) = 7.53$, $p = .481$. In other words, although conversations clearly differed in how much time was devoted to different topics, we did not observe a reliable difference in topic-speaking-time allocation between task-congruent and task-incongruent observations under the primary definition of incongruence.

Nevertheless, Figure 3 highlights that participants devoted substantial conversational space to relational and personal topics, indicating that even in a strongly task-framed setting, interaction was shaped by more than the task objective alone. This is broadly consistent with prior work showing that partner preferences are influenced by task affordances [22]. At the same time, the persistence of topic diversity suggests that task affordances do not eliminate the broader range of informational and relational goals that participants may pursue in conversation.

To complement the topic-distribution analysis, we conducted additional exploratory analyses of conversational behavior at the topic level. We examined topic initiation, topic strength, and topic initiator dominance. Topic initiation, defined as how often speakers introduced new topics within a conversation, was descriptively highly asymmetric: participants who made congruent decisions initiated an average of 94.0% of topics (4.57 topics per conversation on average), and those who made incongruent observations initiated an average of 6.0% of topics (0.31 topics per conversation on average). By contrast, topic strength (Figure 4), defined as how long a topic continued after initiation in terms of number of turns, and topic initiator dominance (Figure 5), defined as the proportion of speaking time within a topic contributed by the speaker who initiated it, did not show clear descriptive differences between congruent and incongruent observations.

*3.4.3 Testing Sensitivity of Task Incongruence Definition.* To assess whether our findings depended on how *Task Incongruence* was operationalized, we conducted a robustness analysis. In the primary analysis, task-incongruent decisions were defined as observations with |studentized residual| > 2 from the linear relationship between change in perceived competence and change in partner selection strength. We re-ran the same topic-distribution analysis using two alternative definitions: 1) a *quarter-based* definition, incongruent decisions were defined as instances where selection strength changed positively while perceived competence changed negatively or vice versa (79 incongruent decisions), and 2) rank-based definition: incongruent decisions were defined as when the change in a partner's perceived competence ranking contradicted

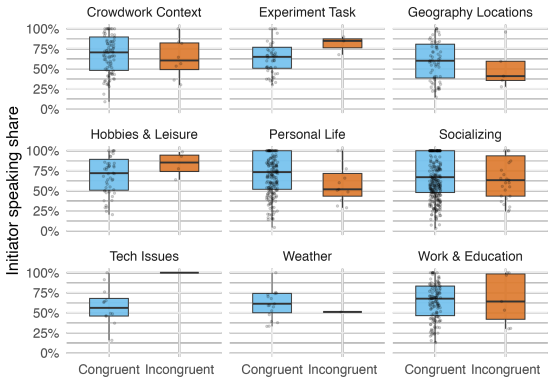



**Figure 5: Exploratory analysis of topic initiator dominance across topic categories. Panels show the distribution of the initiator's speaking-time share within each topic episode, separated by congruency.**

the change in their selection strength ranking among all conversation partners of the decision maker (e.g., a partner's competence rank increased post-conversation, but their selection strength rank decreased), resulting in 51 incongruent decisions.

The descriptive topic distributions varied somewhat across the three definitions (see Appendix C). Statistically, the pattern was partially robust. Across all three specifications, topic category remained a significant predictor of speaking-time allocation (Quartered: $\chi^2(8) = 374.80$, $p < .001$; Rank: $\chi^2(8) = 405.14$, $p < .001$; Studentized: $\chi^2(8) = 418.97$, $p < .001$), indicating that speakers consistently devoted different amounts of time to different topics. Likewise, the main effect of congruency was not significant under any definition (Quartered: $\chi^2(1) = 0.56$, $p = .453$; Rank: $\chi^2(1) = 1.52$, $p = .218$; Studentized: $\chi^2(1) = 1.90$, $p = .168$).

The interaction between topic category and congruency, however, differed across operationalizations. While it was not significant under the quartered definition ($\chi^2(8) = 5.97$, $p = .650$) or the primary studentized definition ($\chi^2(8) = 7.53$, $p = .481$), it was significant under the rank-based definition ($\chi^2(8) = 20.13$, $p = .010$). This suggests that evidence for a difference in topic-speaking-time allocation between congruent and incongruent observations can be sensitive to how decision incongruence is defined.

For the remainder of the paper, we retain the studentized-residual definition as our primary operationalization as it provides the most theoretically precise operationalization of decision incongruence. At the same time, the mixed robustness results also suggest that behavioral manifestations of task incongruence are not fully invariant across operationalizations, further motivating the need for task-aware systems that remain sensitive to uncertainty in how users' latent goals can be inferred.

## 4 Discussion

### 4.1 Task Incongruence in Social Choices and Behaviors

In this empirical investigation, we took a first step towards exploring the misalignment between assumed goals in social interactions and externally observed behaviors in a partner selection setting, where task instructions clearly suggest that the task goal is to choose a competent partner.

Our results suggest that while the majority of partner decisions are consistent with the task requirement (i.e., picking a competent partner), some participants still make task-incongruent choices. Specifically, even when they perceive a partner to be more competent after having a conversation, their preference for that partner still decreases, suggesting that the assumed task goal does not fully capture the goals participants may actually pursue in social interaction. Prior work on this setting also shows selection is driven by both competence and warmth [25, 27], indicating a relational dimension alongside the task goal.

The relatively low rate of incongruent decisions is expected in this highly structured and financially incentivized task setting, where most participants should be aligned with the apparent task goal. We also do not rule out that some portion of these incongruent cases reflects noise or other unobserved factors. Nevertheless, even a modest number of deviations is informative: it shows that a task-goal-only account is insufficient to explain all observed decisions, and therefore cannot be the only basis for modeling users in systems for social reflection.

To gain further insights, we investigated the conversation topics of participants in interactions related to the partner selection task. Under our primary studentized-residual definition, we did not observe a difference in topic-speaking-time allocation between congruent and incongruent observations. At the same time, the descriptive topic ecology showed that conversations were not purely task-oriented: participants still devoted substantial time to relational and personal topics, suggesting that even in a strongly task-framed setting, conversation topics reflect a broader mix of informational and relational goals than the task objective alone.

Additionally, we conducted exploratory analyses of conversational behavior beyond topic prevalence. We examined topic initiation, topic strength, and topic initiator dominance. Topic initiation was highly asymmetric. By contrast, topic strength and topic initiator dominance did not show clear descriptive differences between congruent and incongruent observations. We interpreted these patterns cautiously given the imbalance between congruent and incongruent observations. It's also worth mentioning that identifying the conversational function of utterances is itself non-trivial. The moderate agreement between LLM annotations and human labels underscores that task-aware systems will need richer models of context and user goals than simple topic detection alone.

Finally, it is worth considering how these dynamics might shift in different contexts. In strongly externally driven tasks, the cost of task-incongruent conversation could be high, potentially leading to more task-oriented behaviors. However, in complex collaborations where long-term relationships and trust are crucial, the seemingly task-incongruent prioritization of relationship-building may actually be necessary rather than a deviation from the task.

### 4.2 Implications for Multimodal User Modeling in TSR Systems

Previous work has suggested that TSR systems in the social domain often operate under the assumption where users have a single, predefined goal [20]. However, even in a clearly structured task setting, assuming a single task goal may fail to capture the individual differences underlying task-incongruent choices and behaviors. This can limit TSR systems, as they may encourage users to adjust their actions in ways that move them away from goals that are jointly optimal for the task and for themselves.

To address this, we propose Task-Incongruence Aware (TIA) systems that treat user goals as dynamic and only partially observable, and that support reflection on whether apparent deviations from the assumed task may reflect error, uncertainty, or the pursuit of other equally valid goals. By leveraging the topic diversity and granularity demonstrated in our investigation, TIA systems could provide a mirror that visualizes the balance between social relationship building and instrumental work in interactions.

In addition, task incongruence may also be expressed in multimodal behavioral signals beyond conversational topics, such as gaze or other non-verbal cues. This opens an important direction for future multimodal user modeling in TSR systems. Importantly, such systems may not need to know in advance exactly what alternative goal a user is pursuing. It may already be useful to detect that the user's behavior does not appear aligned with the assumed task goal, and to suggest this mismatch as a prompt for reflection. This would allow users to consider, for example, whether time spent talking about *Personal Life* is a necessary relationship-building step or a distraction, helping them clarify their own goals rather than simply conform to the system's assumptions about the task.

## 5 Limitations and Future Work

There are several limitations we would like to acknowledge in this work. Our automated annotation pipeline reached moderate agreement with human-consensus labels, while the human annotators reached substantial agreement, indicating that the LLM may be less sensitive to certain practical nuances that humans were able to distinguish. Therefore, while our current approach relied on zero-shot prompting, future work should leverage few-shot learning by inserting annotated examples into the prompt to better guide the LLM. Such fine-tuning may improve the LLM's sensitivity to the topics' differences, potentially bridging the gap between human and automated annotators.

In addition, there is limited granularity in these conversation topics. We believe that further analyzing more nuanced subtopic descriptions and deeper analyses of the dynamics of topic changes could provide more insights into how individual goals shape conversational behaviors and partner selection decisions. Finally, as our findings come from a single dataset and one incentivized task, future work should test whether task incongruence generalizes across other social contexts and collaborative tasks.

## 6 Conclusion

By showing that participants sometimes exhibit task-incongruent choices and behaviors that may reflect personal or relational goals beyond the task goal, our work highlights the multifaceted and complex goals that drive human behaviors and decisions in social interactions–even in tasks with clearly defined objectives. These observations suggest that TSR systems in the social domain should move beyond optimizing for a single assumed goal. Instead, having the capability of Task-Incongruence Awareness (TIA) will allow TSR systems to potentially recognize and support multiple, even sometimes divergent, user goals. By doing so, such systems can provide more meaningful reflection opportunities, helping users understand the relationship between their behaviors and their goals in different social contexts.

## 7 Data and Code Availability

The full coding scheme development and automated annotation pipeline implementation are available at: link.

## 8 Usage of Generative AI Statement

Parts of the data processing and analysis code were generated with the assistance of a generative AI tool. In compliance with data privacy protocols, we employed a local instance of Llama 3 for conversation topic extraction. No data was used to train external models, nor was any information shared with model providers. The authors reviewed, modified, and validated all code outputs. In addition, generative AI was employed to refine or rewrite the authors' original text to enhance clarity.

## 9 Safe and Responsible Innovation Statement

Our findings suggest that TSR systems should not infer users' goals from a single assumed task objective, as doing so may misclassify legitimate personal or relational goals as error and provide misleading feedback. However, as the development and deployment of our envisioned TIA systems involve collecting multimodal user data and providing feedback to users, it is crucial to preserve user agency, protect user privacy, and avoid using inferred incongruence for high-stakes automated evaluation.

## 10 Author Contribution

We follow the CRediT (Contributor Roles Taxonomy) [4] to recognize individual author contributions below. Chenxu Hao: conceptualization, methodology, formal analysis, investigation, data curation, writing - original draft, writing - review & editing, visualization, project administration. Tiffany Matej Hrkalovic: conceptualization, methodology, writing - review & editing. Bernd Dudzik: conceptualization, methodology, writing - original draft, writing - review & editing, project administration, supervision. Hayley Hung: conceptualization, methodology, writing - original draft, writing - review & editing, funding acquisition, supervision, project administration.

## Acknowledgments

This research was (partially) funded by the Hybrid Intelligence Centre, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, grant number 024.004.022. This research was partially funded by the ERC funded Project Nonverbal Multimodal Multisensor Social Intention Modelling in Complex Conversational Scenes : Illuminating Novel Insights for Machine Perception (NEON).

## References


[1] Afxentis Afxentiou, Peter Frühmann, Maria Kyriakidou, Maria Patsarika, Daina Gudoniene, Andrius Paulauskas, Alicia García-Holgado, and Francisco José García-Peñalvo. 2022. A case study on gaming implementation for social inclusion and civic participation. In *International Conference on Information and Software Technologies*. Springer, 305–314.

[2] AI@Meta. 2024. Llama 3 Model Card. (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[3] Saliha Akbas, Kemal Kuscu, Asim Evren Yantac, Gizem Erdem, Sinem Semsioglu, Onur Gurkan, Asli Günay, Ali Vatansever, and Terry Eskenazi. 2021. Qualitative dimensions of technology-mediated reflective learning: The case of VR experience of psychosis. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*. 1–10.

[4] Liz Allen, Alison O'Connell, and Veronique Kiermer. 2019. How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing* 32, 1 (2019), 71–74.

[5] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747* (2023).

[6] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. 2009. Action understanding as inverse planning. *Cognition* 113, 3 (2009), 329–349.

[7] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software* 67 (2015), 1–48.

[8] Eric PS Baumer. 2015. Reflective informatics: conceptual dimensions for designing technologies of reflection. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 585–594.

[9] Vanessa Begemann, Lisa Handke, and Nale Lehmann-Willenbrock. 2024. Enabling and constraining factors of remote informal communication: a socio-technical systems perspective. *Journal of Computer-Mediated Communication* 29, 5 (2024), zmae008.

[10] Marit Bentvelzen, Paweł W Woźniak, Pia SF Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting reflection in hci: Four design resources for technologies that support reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.

[11] Harrison Chase. 2022. *LangChain*. https://github.com/langchain-ai/langchain

[12] Carsten KW De Dreu and Laurie R Weingart. 2003. Task versus relationship conflict, team performance, and team member satisfaction: a meta-analysis. *Journal of applied Psychology* 88, 4 (2003), 741.

[13] Patrik Dokoupil, Ludovico Boratto, Ladislav Peska, et al. 2024. Comparing user interfaces for customizing multi-objective recommender systems. In *CEUR WORKSHOP PROCEEDINGS*, Vol. 3815. CEUR-WS, 15–28.

[14] Jeremy A Dressler, Beth A Ryder, Michael Connolly, Megan Dias Blais, Thomas J Miner, and David T Harrington. 2018. "Tweet"-format writing is an effective tool for medical student reflection. *Journal of Surgical Education* 75, 5 (2018), 1206–1210.

[15] Robin IM Dunbar, Anna Marriott, and Neil DC Duncan. 1997. Human conversational behavior. *Human nature* 8, 3 (1997), 231–246.

[16] Rowanne Fleck and Geraldine Fitzpatrick. 2010. Reflecting on reflection: framing a design landscape. In *Proceedings of the 22nd conference of the computer-human interaction special interest group of australia on computer-human interaction*. 216–223.

[17] Alexia Galati, Rick Dale, Camila Alviar, and Moreno I Coco. 2026. Task goals constrain the alignment in eye-movements and speech during interpersonal coordination. *Journal of Memory and Language* 146 (2026), 104691.

[18] Mitsuhiro Goto, Tatsuya Ishino, Keisuke Inazawa, Narimune Matsumura, Tadashi Nunobiki, and Akihiro Kashihara. 2019. Authoring robot presentation for promoting reflection on presentation scenario. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 660–661.

[19] Rebecca Grossman, Kevin Nolan, Zachary Rosch, David Mazer, and Eduardo Salas. 2022. The team cohesion-performance relationship: A meta-analysis exploring measurement approaches and the changing team landscape. *Organizational Psychology Review* 12, 2 (2022), 181–238.

[20] Chenxu Hao, Tiffany Matej Hrkalovic, Daniel Balliet, Hayley Hung, and Bernd Dudzik. 2025. Technologies Supporting Self-Reflection on Social Interactions: A Systematic Review. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 1354–1365.

[21] Tiffany Matej Hrkalovic, Bernd Dudzik, Daniel Balliet, and Hayley Hung. 2025. PARSEL: a Multimodal Dataset for Modeling Decision-Making Processes Involved in Selecting Partners for Joint Tasks. *IEEE Transactions on Affective Computing* (2025), 1–18. doi:10.1109/TAFFC.2025.3600687

[22] Tiffany Matej Hrkalovic, Aria Li, Magnus Bopp, Yingling Li, and Daniel Balliet. 2025. Task affordances affect partner preferences. *Journal of Experimental Social Psychology* 119 (2025), 104751.

[23] Zeyuan Huang, Qiang He, Kevin Maher, Xiaoming Deng, Yu-Kun Lai, Cuixia Ma, Sheng-Feng Qin, Yong-Jin Liu, and Hongan Wang. 2023. SpeechMirror: A multimodal visual analytics system for personalized reflection of online public speaking effectiveness. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2023), 606–616.

[24] Dietmar Jannach and Himan Abdollahpouri. 2023. A survey on multi-objective recommender systems. *Frontiers in big Data* 6 (2023), 1157899.

[25] Vasiliki Kentrou, Jacek Buczny, Tiffany Matej Hrkalovic, Bernd Dudzik, Daniel Balliet, and Reinout E de Vries. 2026. Partners among Strangers: A social Relations perspective on personality and collaborative partner preferences in first encounters. *Journal of Research in Personality* (2026), 104724.

[26] Sung-In Kim, So-youn Jang, Taewan Kim, Bogoan Kim, Dayoung Jeong, Taehyung Noh, Mingon Jeong, Kaely Hall, Meelim Kim, Hee Jeong Yoo, et al. 2024. Promoting self-efficacy of individuals with autism in practicing social skills in the workplace using virtual reality and physiological sensors: Mixed methods study. *JMIR Formative Research* 8 (2024), e52157.

[27] Tiffany Matej Hrkalovic, Bernd Dudzik, Hayley Hung, and Daniel Balliet. 2025. Partner perceptions during brief online interactions shape partner selection and cooperation. *PloS one* 20, 4 (2025), e0318137.

[28] Antonija Mitrovic, Vania Dimitrova, Amali Weerasinghe, and Lydia Lau. 2016. Reflective experiential learning: Using active video watching for soft skills training. In *Proceedings of the 24th international conference on computers in education*. Asia-Pacific Society for Computers in Education.

[29] Ollama. 2023. *Ollama: Large Language Model Runner*. https://ollama.com

[30] Viktoria Pammer-Schindler and Michael Prilla. 2021. The reflection object: An activity-theory informed concept for designing for reflection. *Interacting with Computers* 33, 3 (2021), 295–310.

[31] Ava Elizabeth Scott, Lev Tankelevitch, and Sean Rintel. 2024. Mental models of meeting goals: supporting intentionality in meeting technologies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

[32] Herbert Alexander Simon. 1997. *Models of bounded rationality: Empirically grounded economic reason*. Vol. 3. MIT press.

[33] Carmen Tabernero, M Jose Chambel, Luis Curral, and Jose M Arana. 2009. The role of task-oriented versus relationship-oriented leadership on normative contract and group performance. *Social Behavior and Personality: an international journal* 37, 10 (2009), 1391–1404.

[34] Emma Templeton, Luke Chang, and Thalia Wheatley. 2024. Conversational launch pads: Strangers start their conversations with topics that lead to many other topics. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 46.

[35] Huimin Wei, Simeng Zhang, Wenran Qu, Mengqi Liu, Zeping Yan, and Xiaorong Luan. 2024. The association between team cohesion and performance: A network analysis of nurses. *Nursing & Health Sciences* 26, 1 (2024), e13089.

[36] Michael Yeomans, Maurice E Schweitzer, and Alison Wood Brooks. 2022. The conversational circumplex: Identifying, prioritizing, and pursuing informational and relational motives in conversation. *Current Opinion in Psychology* 44 (2022), 293–302.

[37] Hangyu Zhou, Yuichiro Fujimoto, Masayuki Kanbara, and Hirokazu Kato. 2021. Virtual reality as a reflection technique for public speaking training. *Applied Sciences* 11, 9 (2021), 3988.