TU Delft

Delft University of Technology

PARSEL

a Multimodal Dataset for Modeling Decision-Making Processes Involved in Selecting Partners for Joint Tasks

Hrkalovic, Tiffany Matej; Dudzik, Bernd; Balliet, Daniel; Hung, Hayley

DOI
10.1109/TAFFC.2025.3600687

Publication date
2025

Document Version
Final published version

Published in
IEEE Transactions on Affective Computing

Citation (APA)
Hrkalovic, T. M., Dudzik, B., Balliet, D., & Hung, H. (2025). PARSEL: a Multimodal Dataset for Modeling Decision-Making Processes Involved in Selecting Partners for Joint Tasks. *IEEE Transactions on Affective Computing*, *16*(4), 3481-3498. https://doi.org/10.1109/TAFFC.2025.3600687

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Green Open Access added to TU Delft Institutional Repository
as part of the Taverne amendment.

More information about this copyright law amendment
can be found at https://www.openaccess.nl.

Otherwise as indicated in the copyright section:
the publisher is the copyright holder of this work and the
author uses the Dutch legislation to make this work public.

PARSEL: A Multimodal Dataset for Modeling Decision-Making Processes Involved in Selecting Partners for Joint Tasks

Tiffany Matej Hrkalovic, Bernd Dudzik, *Member, IEEE*, Daniel Balliet, and Hayley Hung, *Member, IEEE*

***Abstract*—How people evaluate, select, and engage with others in cooperative settings significantly impacts their well-being, happiness, and success. However, navigating these processes is complex. Equipping systems with the ability to recognize, interpret, and even engage during such socio-cognitive processes can increase their potential to support humans in these socio-cognitive processes and be more successful in adjusting to the social environment they are embedded in (e.g., understanding human preferences and attitudes), leading to better quality interactions and decision-making for future partners. Yet, the developments of such systems depend on available datasets. However, based on our knowledge, no dataset exists that can be used to model partner selection for joint tasks. To support research focused on creating such intelligent systems, we introduce the PARSEL dataset – a comprehensive corpus of dyadic interactions designed for computational modeling of *PAR*tner *SEL*ection processes and collaborative behavior. In total, 297 participants took part in the datasets. The dataset contains measurements of partner selection decisions over three different stages, as well as factors that may influence partner selection in the context of (online) social interactions. It includes audiovisual recordings that offer fine-grained behavioral cues used during these interactions, self-reported traits, and reported perceptions of person-, situation-, and team-specific phenomena. By providing this resource, we aim to foster advancements in computational methods that can effectively model and augment socio-cognitive processes, contributing to socially aware intelligent systems and enhanced human-system interactions.**

***Index Terms*—Cooperation/collaboration, partner selection, social perceptions, social signal processing.**

Received 15 April 2025; revised 15 July 2025; accepted 11 August 2025. Date of publication 19 August 2025; date of current version 3 December 2025. The work of Daniel Balliet was supported by ERC Consolidator under Grant 864519. This work was supported by the Hybrid Intelligence Center under Grant 024.004.022. *(Tiffany Matej Hrkalovic and Bernd Dudzik contributed equally to this work.) (Corresponding author: Tiffany Matej Hrkalovic.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Ethical Board of the VU Amsterdam under Application No. VCWE-2021.168.

Tiffany Matej Hrkalovic is with the Amsterdam Collaboration Lab (ACL), Vrije Universiteit Amsterdam, 1081 BT Amsterdam, The Netherlands, and also with the Department of Intelligent Systems (INSY), Delft University of Technology, 2628XE Delft, The Netherlands (e-mail: tmatejhrkalovic@tudelft.nl).

Bernd Dudzik and Hayley Hung are with the Department of Intelligent Systems (INSY), Delft University of Technology, 2628XE Delft, The Netherlands (e-mail: b.j.w.dudzik@tudelft.nl; h.hung@tudelft.nl).

Daniel Balliet is with the Amsterdam Collaboration Lab (ACL), Vrije Universiteit Amsterdam, 1081 BT Amsterdam, The Netherlands (e-mail: d.balliet@vu.nl).

This article has supplementary downloadable material available at https://doi.org/10.1109/TAFFC.2025.3600687, provided by the authors.

Digital Object Identifier 10.1109/TAFFC.2025.3600687

I. Introduction

HUMANS participate in a myriad of cooperative and collaborative interactions daily. These interactions, when successful, help humans thrive by forming new and maintaining old relationships [1]. For these interactions, selecting suitable partners is crucial, as the choice of a partner can have a significant role in determining the short-term or long-term success of such interactions [2], [3]. Despite the importance of partner selection, less research is oriented on understanding *how* people select partners for cooperative and collaborative interactions.

Choosing suitable partners for cooperative and collaborative interactions is a crucial but challenging decision-making process. This process involves multiple *sequential* steps where individuals must first identify and evaluate others, to then select the most suitable partners [3]. For instance, when interacting with others, individuals benefit from forming perceptions of their partner's traits (*"Are they trustworthy?"*), assessing the dynamics of their interaction (*"Are we getting along well together?"*), considering the requirements of the future interaction or task (*"For what type of task or interaction do I need a partner?"*), their affective state and many more [4]. These perceptions are often based on limited social cues, such as verbal and non-verbal behaviors, that lead to very rapid but sometimes erroneous perceptions of potential partners [5], [6]. Despite the potential for error, existing research shows that individuals rely on a multitude of these perceptions to guide their partner selection and subsequent actions [3]. However, inaccurate perceptions can hinder successful collaboration by leading to missed opportunities. For example, if someone mistakenly perceives a potential partner as unkind or angry, they may be hesitant to engage with them, potentially overlooking a beneficial collaboration.

Understanding how people select partners offers promise for improving these decisions. Improvement of such decisions and, consequently, future human interactions, can be done via human interventions or intelligent systems support (see [7]). For instance, if literature proposes that people tend to overestimate the role of people's trustworthiness and kindness when selecting partners [8], this information might help the system to nudge the user in the right direction by making other cues more salient, such as emphasizing that the task requires a competent partner. These solutions can help promote self-reflection and mindfulness when evaluating and selecting partners. For example, giving individuals insight into how their (non-)verbal behavior

1949-3045 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

is interpreted and evaluated by others can inform their own and others' partner selection. Making this information explicit can enhance individuals' socio-cognitive skills and improve their future interactions [9]. While human interventions have been effective in enhancing decision-making in various areas of social interaction, support via intelligent systems offers the advantage of providing real-time, consistent, and adaptable support [10]. However, delivering such support requires systems to master multiple sub-tasks, including inferring human perceptions and preferences for future partners.

A successful design of such intelligent systems rests on available datasets that provide data for computational modeling and training of such systems [11]. Modeling can focus on tasks such as inferring how people perceive each other or their interaction, and most importantly, their preferences. All of these tasks could help make an intelligent system that can support human behavior and partner selection in social interaction. While numerous datasets exist for studying socio-cognitive phenomena, such as personality perception [12], [13], none focus on the sequential decision-making stages involved in selecting partners for collaboration, let alone how these decisions are coupled with tangible outcomes such as team performance. Developing effective methods to enhance the quality of interpersonal interactions is a significant topic in affective computing [7], making it a valuable area for further investigation.

Therefore, in this paper, we present *PARSEL* – a novel multimodal dataset collected to design a dataset for modeling the successive phases of the decision-making process involved in selecting suitable cooperation and collaboration partners. Thus, this dataset measures diverse perceptions people have of each other, the situation and the interaction, and partner selection itself. We recruited 297 participants in groups of up to six strangers and instructed them to find a suitable partner among their fellow group members to solve a cooperative and collaborative task following a round-robin design. Thus, there was an underlying goal of the interactions to select a suitable partner for two joint tasks. To do so, they first observed individual photographs of all other participants, engaged in a series of pair-wise conversations with each other, and made partner selection choices. Finally, they engaged in dyadic cooperative and collaborative tasks with all other participants. The final dataset contains:

1) audiovisual recordings of both initial conversational and collaborative task interactions,
2) a broad variety of self-reports measured with validated questionnaires describing relevant aspects of participants' background (e.g., HEXACO personality),
3) first-person perceptions and reports of decision-making outcomes (i.e., Person Perceptions (*"How trustworthy/skillful is this person?"*), Situation Perceptions (*"Our preferred outcomes in this situation are conflicting."*), and Team perceptions, and finally
4) measures of cooperative and collaborative behavior and performance (e.g. *money exchanged in a cooperative task* and *correctness of answers on a question from a Family Feud*).

This dataset was utilized in previous research to model person perceptions of warmth and competence and investigate the role of these perceptions in predicting their role in partner selection (see [9]). Nevertheless, given the richness of the dataset, we believe it can support other research tasks, such as modeling conversation and collaboration dynamics, and further investigate the role of other aspects of interactions to model partner selection (e.g., synchrony of non-verbal behaviors) (see Potential Uses). Motivated by this, this paper makes the following contributions:

- *Presentation of PARSEL:* We present and document the design and content of PARSEL.
- *Dataset Description and Validity:* We provide statistical descriptions of the data.
- *Demonstration of Dataset Usefulness for Partner Selection Modelling:* We demonstrate the usefulness of the dataset for modeling partner selection decisions and other social phenomena related to social interactions.
- *Uses for Future Research:* We discuss how *PARSEL* can be used for future research focused on computational modeling of social phenomena, such as affective and conversational expectations.

II. Motivational Axes for Constructing *PARSEL*

In this section, we outline the key motivational axes and previous key research that guided the design of PARSEL. PARSEL was structured around three *motivational axes*, each broken down into *design principles* that shaped our methodological choices (see Scenario and Experimental Setup).

Axis 1: Comprehensive Measurement of Partner Selection

Partner selection is sequential and multi-criteria, where individuals engage in several evaluative processes [14], consider various factors, and strategically choose partners to fulfill their task or relational goals. These decisions have measurable consequences on future behavior. To ensure a comprehensive approach, PARSEL follows two key design principles that acknowledge the aforementioned qualities.

A1.P1*: Consider the Sequential Nature of Partner Selection and Evaluation Processes*

Partner selection occurs through multiple phases, beginning with gathering information about potential partners and culminating in an explicit or implicit decision to select a partner for future collaboration [3]. The initial phase typically involves identifying potential partners and gaining insights about them through conversational interactions [15]. These perceptions and evaluations serve as the foundation for assessing partner suitability and should influence people's partner selection [14].

However, the experience of any given interaction, and therefore its behavior, is shaped not only by immediate *person perceptions* but also by how people cognitively assess social interactions [16]. Similarly, findings indicate that salient *Remembered Moments* play a crucial role in evaluating the overall experience of situations [17]. Remembered information can strongly influence future behavior (e.g., whether to expose oneself to a particular situation again [18]). Consequently, *which* moments from a conversation an individual does remember and *how* exactly these are evaluated in retrospect may be crucial

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

information for studying and modeling partner selections. Except for these, prior research also indicated that individuals' *Expectations* of how an interaction will play out before they engage in it may influence their corresponding behavior and perceptions of experience (e.g., in terms of surprise [19]).

Thus, capturing these phases enables us to model the antecedents that naturally influence partner selection in social situations and people's behavior in future tasks. Failing to address this sequential nature would limit the applicability of PARSEL in understanding partner selection dynamics. Thus, to model this complexity, we designed a dataset that captures the sequential development of partner selection across its various stages, enabling us to understand how individuals move from initial interactions to final decisions.

A1.P2*: Include Different Components Informing Partner Selection*

Except for the different temporal phases of decision-making, partner selection is also shaped by multiple contextual factors, including person-specific (e.g., personality, preferences, affective state) [20], situation-specific (e.g., power dynamics, task requirements) [21], and partner-specific factors (e.g., non-verbal behavior, attitudes) [15]. These factors collectively make partner selection a complex decision-making process. For instance, it is known that perceiving a situation as high on the conflict of interest may reduce the willingness of individuals to cooperate ([22]), but also potentially reduce the likelihood of being selected as a partner.

However, usually, existing datasets do not integrate all these influencing factors. Capturing this data is essential for a deeper understanding of how people choose partners and for developing intelligent systems that can support decision-making in social interactions, as they need to navigate and account for all of these factors.

To address this, PARSEL measures key factors affecting partner selection across two tasks requiring different partner profiles. Additionally, we introduce conditions that favor selecting either warm or competent partners to induce changes in situational affordances (i.e., barriers or opportunities of a situation for expression of certain traits or behaviors [23] Scenario and Experimental Setup).

Axis 2: Maximizing Ecological Validity

To create intelligent systems that support human users in everyday social interactions, it is essential to train models on data that adequately represents real-life situations. Failing to do so can have negative consequences on model performance during real-life deployment. Here, we present two design principles that have guided the design of PARSEL and are aligned with this motivation.

A2.P1*: Use In-the-Wild Recording Conditions*

Inspired by motivational rationales from other datasets (see [24]), an important design principle was to go beyond the highly controlled video recording conditions provided in laboratory settings. Laboratory settings are useful for providing high-quality audio and video input, but they underestimate the variance of recording conditions in real-life settings and potentially convey unrealistic expectations. To create intelligent systems that are robust to changes in recording conditions, such as uncontrolled lighting, and many more, such systems need to be trained on data containing such variations.

Thus, another goal of PARSEL was to capture audiovisual data containing natural variance in recording conditions.

A2.P2*: Consider The Influence of Personal Background*

Various person-specific factors influence social interactions and decision-making, including personality traits (e.g., extraversion, agreeableness) [25], trust [26], social anxiety [27], and social value orientation [28]. For instance, extroverts tend to be more at ease in social settings and are more likely to self-disclose [29]. Socially anxious people tend to behave differently during social interactions [30].

Controlling for these factors is essential to ensure the dataset representativeness and generalizability, but also better performance in predicting partner selection. A dataset dominated by extroverted individuals, for example, may not reflect broader social patterns. Moreover, models should account for these variations to enhance contextual awareness and robustness.

To address this, PARSEL systematically measures relevant personal background variables for partner selection and person perception, as detailed in Scenario and Experimental Setup.

Axis 3: Facilitation of Interdisciplinary Research

Modeling social behaviors, such as personal attitudes and partner selection, requires an interdisciplinary approach that bridges Social and Computational Sciences [31]. Social scientists provide insights into human behavior, while computational scientists translate these into models and algorithms.

Interdisciplinary datasets are essential for advancing human-centered intelligent systems [32]. They enable social scientists to refine computational models and help computational scientists develop models that better reflect real-world behaviors. However, such datasets must meet the standards of both fields. For example, modeling facial expressiveness requires machine-readable data [11], while studying social phenomena such as emotions and partner selection requires validated psychological measures.

To maximize interdisciplinary accessibility, PARSEL integrates standards from Social Science (validated measures), Cognitive Science (sequential partner selection design), and Computational Science (multimodal machine-readable data). This allows researchers across disciplines to analyze the same data from different perspectives.

III. Related Work

A wide range of multi-modal datasets exist for modeling various aspects of social interactions, including (non-)verbal behavior, group dynamics [33], [34], [35], [36], [37], rapport [38], impressions [39], personality [13], online conversations [40], and partner selection in speed-dating [41]. However, there is relatively little data collected with the aim of modeling partner selection, with no existing datasets for modeling sequential partner selection for joint undertakings.

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARISON OF PARSEL WITH EXISTING DATASETS

Dataset	Data Collection			First-person evaluations				Sensor Data	Collaborative, Cooperative Setting and Performance	Partner Selection Decision	Personal Background		(Sequential) Design	
Dataset Name	Type	N	Setting	Person perceptions (W/C)	Person perceptions (Personality)	Conversation Perceptions	Team Perceptions	Behavioral			Demographics	Self-reported Traits	Sequence of Interaction Stage	Sequential design of Decision-Making Process
MULTISMO	IP	49	Lab		+	+		+			+	+		
UGI	IP	86	Lab			+	+	+	+		+			
AMI	IP	-	Lab										+	
ICSI	IP	53	Lab								+			
IMPRESSION	SP	54	Lab (Online)	+				+			+			
First Impressions dataset	SP	-	Non-interactive (Online)					+			+			
NoXi	IP	87	Lab					+			+	+		
CANDOR	SP	1456	Online	+	+			+		+	+	+		
MatchNMingle	IP	92	Face-to-Face					+		+	+	+	+	+
PARSEL	SP	297	Online	+	+	+	+	+	+	+	+	+	+	+

Nevertheless, given the complexity of partner selection and socio-cognitive processes that relate to it, in this section, we provide an overview of existing datasets that were either used for motivation when designing PARSEL or were designed for modeling social phenomena and communication behavior which can also be modeled using PARSEL, such as impression formation and collaborative dynamics (see Potential Uses). For example, despite PARSEL being designed mainly for the comprehensive modeling of partner selection, PARSEL can also be used to model other social phenomena, such as team perceptions or situation perceptions, regardless of partner selection.

To make a distinction between the original goals of each dataset, we make a distinction and focus on two types of datasets: (a) Interaction Process (IP) datasets, which focus more on modeling the behavioral patterns in interactions, such as turn-taking dynamics, and (b) Social Cognition Phenomena (SP) datasets, which focus more on modeling social-cognitive phenomena, such as perception of warmth and competence. Some datasets overlap in the goal of providing data for modeling both interaction and cognition.

A. Datasets Measuring Interaction Processes (IP)

Most efforts in this cluster rest on different aspects of social interaction, to capture behavioral dynamics while solving collaborative tasks (*MULTISMO* and *UGI*) [37], [38], with a minority of the datasets capturing contexts of business meetings (*AMI* and *ICSI*) [34], [35] or informal conversations (*CANDOR and MatchNMingle*) [40], [41], but none comprehensively capture partner selection in cooperative and collaborative tasks.

For example, both *MULTISMO* [37] and *UGI* [38] capture a group of individuals solving a collaborative task. This collaborative task is the same as the collaborative task used in PARSEL. However, MULTISMO and UGI fall short in the efforts of measuring partner selection, as they do not measure any relevant perceptions or partner selection decisions that are needed to model the sequential process of partner selection (see Table I). Similarly, AMI and ICSI examine behaviors in business meetings but lack data on partner selection decisions or relevant social perceptions. On the other hand, NoXi contains records of dyadic interactions and engagement cues but only includes sensory data and third-party annotations, missing direct measures of partner evaluation (see Table I). However, NoXi includes behavioral cues (e.g., eye-gaze, backchannels) during information exchange and provides continuous annotations of participant engagement, enabling modeling human behavior during dyadic interactions [42].

The last subset of datasets from this group includes two datasets: *CANDOR* [40] and *MatchNMingle* [41]. The two datasets both contain measures of partner selection, but for different contexts than PARSEL. For instance, CANDOR has the most overlap with PARSEL. It captures dyadic online conversations, measuring warmth, competence, and the desire to talk again—potential proxies for partner selection [43]. However, it does not model the sequential decision-making process or tie selections to a specific future task. Finally, CANDOR measures partner selection hypothetically; there is no specific future task to which the decision is tied. On the other hand, the MatchNMingle dataset [41] focuses on romantic partner selection within a speed-dating setting, capturing sequential interactions and their consequences, such as follow-up mingling sessions. While it provides insights into partner selection outcomes, it is designed for studying dating behaviors rather than partner selection for joint tasks. Unlike PARSEL, MatchNMingle was collected to analyze social interaction processes and behaviors in the context of romantic dates, instead of modeling partner selection as a cognitive process.

B. Datasets Measuring Social Cognition Phenomena (SP)

Datasets focused on cognitive processes have also been collected in social interaction contexts. However, here we include datasets that were also collected for modeling and understanding cognitive processes, such as person perceptions, instead of only looking at interaction behaviors and their emerging properties, such as engagement. Currently, two datasets, *IMPRESSIONS* [39] and *First Impressions* [13], exist that can be used for modeling perceptions of personality and warmth, and competence.

The First Impressions dataset consists of YouTube videos of strangers where annotators evaluate the hirability and personality of each person presented in the video. In contrast to PARSEL, the First Impression dataset only includes perceptions by people looking at videos instead of being included in the interaction (i.e., third-person annotations of person perceptions, instead of PARSEL's first-person perceptions; see Table I). Usually, third-person perceptions are devoid of any social interaction. On the other hand, the IMPRESSIONS dataset [39] included perceptions of third- and first-person perceptions of warmth and competence. In the IMPRESSION dataset, participants

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

provided perceptions of warmth and competence during natural interactions with people who were included in the conversation (i.e., first-person perceptions). These annotations were collected continuously during the interaction; it is unclear if this could contaminate spontaneous social behavior due to the additional cognitive load of live self-reporting. Additionally, the IMPRESSION dataset does not include measurements of partner selection or other relevant perceptions of the situation or the team (see Table I).

PARSEL stands apart from the aforementioned datasets by integrating multiple aspects of social interaction. It captures both informal conversations and structured collaborative tasks, allowing researchers to study behavioral consistency across different contexts. Unlike other datasets, it systematically measures partner selection as a step-by-step process tied to real future interactions. Moreover, it combines behavioral and cognitive data, including perceptions of social warmth, competence, and decision-making rationales, making it a more comprehensive resource for modeling partner selection in cooperative settings.

IV. Data Collection Framework

In this section, we outline a detailed description of the design and the entire study procedure.

A. Participants

To recruit participants, we used a well-known online platform for recruiting participants - Prolific. Participant selection was limited to the subset of individuals who were living in the U.K. at the time of data collection, were proficient in English, and had the necessary equipment (i.e., headphones, web camera, bandwidth, laptop/PC) and settings (i.e., quiet room with limited noise and distractions). Throughout the study, participants were instructed to use their laptops or Personal Computers and position them on a stable surface. A stable surface was used to ensure that we had a similar setting for all participants, as well as good visibility of their faces and upper bodies. All participants needed to be older than 18 years and consent to the sharing, recording, and use of their data for research purposes.

B. Scenario and Experimental Setup

The dataset was collected with a study consisting of two parts: 1) the Intake part and 2) the Interaction part.

1) The Intake Part: The Intake phase included a consent form, a schedule availability check, an internet bandwidth test, and a self-report questionnaire. Before proceeding, participants had to consent to data sharing, recording, and research use. To align with Principle A2. P2, all participants were asked to report on their personality, social anxiety, psychopathic traits, trust, prosocial behavior, fluid intelligence, and partner preferences in cooperation (see Section Survey Measures and Materials). These measures were chosen based on prior research linking personality [44], anxiety [45], psychopathy [46], and preferences [47] to social behavior. Additionally, prosociality, benevolence, and trust were assessed due to their influence on cooperation [48]. Participants then scheduled a slot for the Interaction phase. The questionnaire was administered online via Qualtrics.

Between the Intake part and the Interaction part, all participants received an email through the Prolific platform containing a) the instructions for the Interaction part, b) the link they needed to use to access the study, and c) the time and date.

2) The Interaction Part: The Interaction phase occurred one week after the Intake. Participants were grouped (4–6 members) based on their reported availability, with only one group participating at a time. While the initial goal was six-member groups, technical issues and cancellations led to variations in size. Following a round-robin design, each participant interacted with all group members.

To access the Interaction part of the study, participants needed to use a link given to them via email in Prolific, which led them to the landing page. This part combined a Qualtrics survey with custom JavaScript elements for progress monitoring and in-browser video conferencing (for more details, see Section Web application for Data Collection). Upon arrival at the welcome page, participants were instructed to wait to be contacted by one of two available researchers via video call to check the validity of their setup (i.e., a working web camera and a good bandwidth). If a participant's setup was satisfactory, they were forwarded to the first page of the survey, where they waited for other participants to finish with the validation. If a participant's setup was unsatisfactory, they were redirected to a separate study on social relationships, as they had been informed beforehand, but this study is not part of this paper.

After passing the landing page, all participants were briefed on their goal in the Interaction part: selecting suitable partners for a Joint Task. Each group was randomly assigned to one out of two versions of this task: either the *Joint Trust Task* or the *Joint Competence Task*. Half of the groups were required to select partners for a *Joint Trust Task* that emphasized partner warmth (i.e., trustworthiness). The Joint Trust Task was a modified two-player version of the Trust Game [49], assessing trust and reciprocity. Each player received 10 monetary units (MU) and decided how much to send to their partner, with the amount increasing by 20% during the transaction. Cooperation allowed both players to maximize earnings, but success depended on mutual trust. MUs were later converted into real money (1 MU = £0.50) and paid as a bonus (see Appendix for details). The other half of the participants selected partners for the Joint Competence Task, which focused on competence (i.e., intelligence). Unlike the Joint Trust Task, this was a newly developed decision-making challenge created by two of the paper's authors. Participants had aligned interests but could only achieve mutual benefits by solving intelligence-based problems from a validated non-verbal intelligence test. Success depended on problem-solving ability rather than trust (see Appendix for details).

To ensure participants understood the task, they watched an instructional video specific to their assigned version. Afterward, they answered three comprehension questions before proceeding to the next stage.

Impression Stage: After learning about the Joint Task, participants entered the Impression Stage. They first took a photo using their webcam for two purposes: capturing first impressions

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

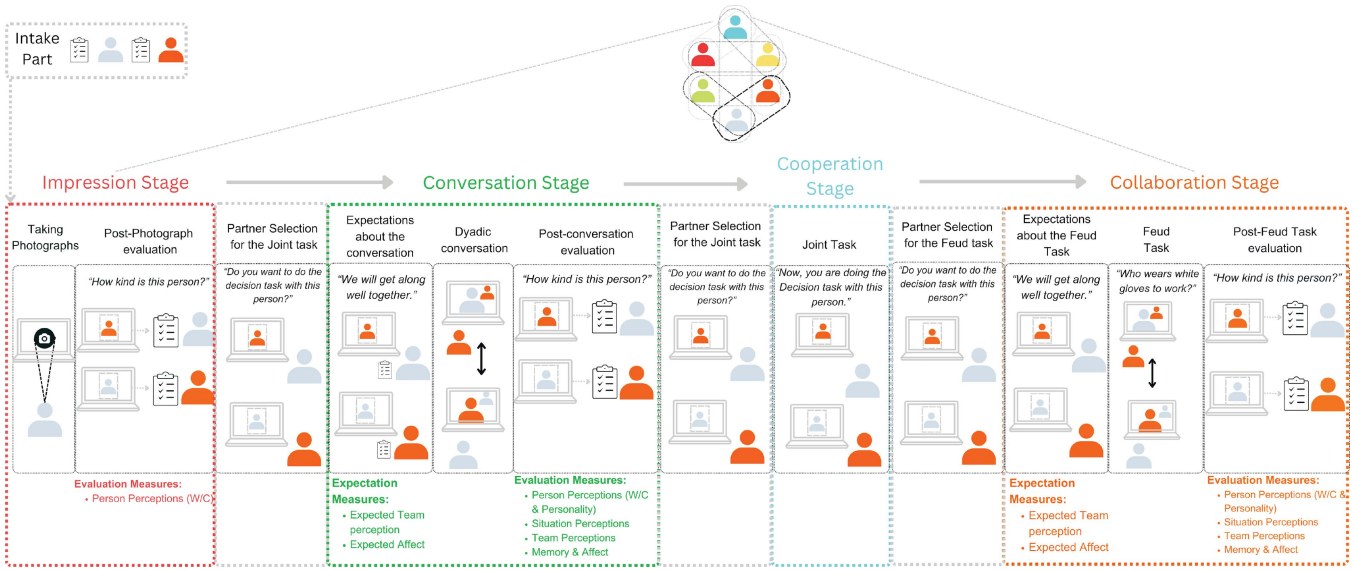

Fig. 1. Overview of the interactive part of PARSEL dataset. The figure visualizes a participant's journey in the interactive part of the dataset. Every participant in the group of 4-6 participants went through all four stages (impression, conversation, cooperation, and collaboration). Here, we only present the journey of two participants (blue and orange in the picture). We used a round-robin design, which means that all participants evaluated photographs from all participants within their group (n-1), were paired, and talked to all participants for the Dyadic conversation and the Feud task.

based solely on appearance and facilitating study navigation. Next, they viewed photos of all group members and rated them on perceived warmth and competence (see Red Box in Fig. 1). They then indicated whether they wanted to partner with each person and, if so, how strongly (see Gray Box in Fig. 1).

Conversation Stage: The Conversation Stage followed the Impression stage. At the start of the Conversation Stage, participants were informed they would have a three-minute, one-on-one video conversation with each group member. Before each conversation, they viewed the other person's photo and estimated how well they would get along (Expected Team Perception) and how the conversation might make them feel (Expected Affect). After setting their expectations, participants had a conversation. Following each interaction, they completed a post-conversation survey to capture their evaluations, including ratings of the other person's warmth and competence (Person Perceptions (W/C)), personality perceptions (Person Perceptions of Personality), situational and conversation dynamics (Situation Perceptions), team dynamics (Team Perceptions), and their affective state (see Survey Measures and Materials for details for details). They also described a memorable event from the conversation (Memory). These measures were based on prior research highlighting the importance of partner preferences for warmth, competence, and personality in partner selection [50]. Additionally, the Situational Interdependence theory was used to capture perceptions of power dynamics, conflict, and mutual dependence in interactions, rather than just objective situational characteristics (e.g., "How demanding is the situation?" [51]). Perceptions of conversational dynamics, such as rapport [52], and team dynamics, such as entitativity and cohesion, were also important for understanding interpersonal liking and future collaboration [53], [54]. Finally, participants were asked to recall the most memorable moment of the conversation, which helped assess how it affected their general perceptions and emotions (see Green Box in Fig. 1, and Principle A1.P1 for motivation).

Guided by another motivation axis to capture realistic recordings of online interactions (see Principle **A2.P1**), we aimed to capture natural variation in terms of recording quality. All video conversations were recorded via participants' web cameras and without specific constraints regarding web camera models or lighting conditions. However, we did ask them to place their computer on a solid surface, showing their head clearly. This process resulted in two separate recordings for one interaction (one for each participant).

Cooperation Stage: Once all conversations and evaluations were completed in the Conversation Stage, participants were asked to select partners for the Cooperation Stage (second Gray Box in Fig. 1). However, regardless of their selections, participants were still paired with every other group member to complete the Joint Task. This approach aligns with one of the aims of the dataset, understanding the consequences of partner selection in terms of its consequences on cooperative behavior. Unlike the live video interactions in the Conversation Stage, no live interaction occurred here. Instead, participants saw a photograph of their partner and completed the task with them, using the photo to identify their partner (see the Blue Box in Fig. 1). After completing the Joint Task with each member of the group, the participants were asked to report their beliefs about how the other person behaved in the task. Importantly, participants were not told how well they performed in the Joint Task to avoid influencing their behavior in the upcoming Collaborative Stage.

Collaboration Stage: After completion of the Joint Task, participants were introduced to an unexpected component of the study: the Feud Task. The Feud Task is a collaborative dyadic assignment introduced only after the Joint Task to avoid confusing participants about which task they were selecting

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

partners for. To separate the partner selection process for the Joint Task from the Feud Task, participants were once again asked to select partners specifically for the Feud Task. The Feud Task was inspired by the television game show Family Feud, which was already used in previous research [37]. In this task, participants were given a question and had to collaborate with another participant to list the three most popular answers to that question. They then ranked these answers from most popular to least popular. The task was completed via a three-minute video chat, and participants engaged in a different question with each member of the group. After each Feud Task, participants filled out an evaluation form assessing their collaboration partner, the interaction, and the situation (see the Orange Box in Fig. 1). The form was similar to the one used in the Conversation Stage, but it also included measures of Entitativity and Task Cohesion as part of Team Perceptions (see Survey Measures and Materials).

To prevent any data loss, at the end of the Interaction part, participants were asked to manually upload their audio-video recordings if the automatic upload had failed. Finally, participants were debriefed about the purpose and design of the study.

C. Survey Measures and Materials

1) The Intake Part Questionnaires: The Intake part questionnaires were used to collect any person-specific components that can contribute to explaining how people behave in social interactions and make decisions (see Principle **A2.P2**). Thus, here we collected self-reports of personality, social anxiety, psychopathy, prosocial behavior, and partner preferences.

- *Demographics* were captured by asking for the basic demographic information, including age, sex, and nationality of the person.
- *Personality* was measured with HEXACO-60 [55]. HEXACO is used to measure six personality dimensions (i.e., Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience) of the HEXACO model. All dimensions are measured on 60 (10×6) items. Participants indicated their agreement with each item on a 5-point Likert scale (1 – "Strongly Disagree" to 5 – "Strongly Agree"). The HEXACO model was used over other personality models (i.e., Big Five), as it measures Honesty-Humility, which has an important role in prosocial behavior [56].
- *Social Interaction Anxiety* was measured with the Social Interaction Anxiety Scale (SIAS-6; [57]). SIAS-6 is a shortened form of the Social Interaction Anxiety Scale. The scale consists of six items on which participants need to report their agreement on a 5-point Likert scale (1 – "Strongly Disagree" to 5 – "Strongly Agree"). Higher values indicated higher levels of anxiety. This measure was specifically used as it was focused on measuring anxiety in the context of social interactions.
- *Psychopathic Personality* was measured with the Psychopathic Personality Traits Scale (PPTS). PPTS is a 20-item measure measuring four dimensions: affective responsiveness, cognitive responsiveness, interpersonal manipulation, and egocentricity. Each dimension consists of five items for which participants need to state their agreement (1 – "Strongly Disagree" to 5 – "Strongly Agree"). Higher scores indicate higher levels of psychopathic personality.
- *Trust* was measured with a scale measuring proposed factors of trust: ability, benevolence, and propensity [26]. Ability was measured with a six-item scale. Propensity was measured with an eight-item scale previously used in [26]. Finally, benevolence was measured with a five-item scale. This measure was used as it adopts a multi-dimensional definition of trust [26].
- *Prosociality* was measured using Social Value Orientation Slider Measures (SVO; [28]), which consists of six items measuring a distributional preference between self and other. In each item, individuals needed to decide how much they wanted to give to the other person and how much to themselves. The overall SVO score is computed based on each item's decision. This measure of prosociality was used as it is a well-established and validated measure of prosocial behavior [58].
- *Partner preferences* were measured with an altered ideal partner design question (see [59]). In this task, participants needed to allocate a budget to eight characteristics (i.e., morality, sociability, competence, emotional stability, open-mindedness, attractiveness, similarity, and consciousness). Each participant was given 100 coins to distribute among these eight characteristics. Higher amounts of coins indicated a higher importance of the characteristic. We used a well-established method of measuring partner preferences [59].
- *Non-verbal intelligence and reasoning* were measured using the University of California Matrix Reasoning task (UCMRT; [60]). The UCMRT consists of 23 matrices, two example problems, and six practice problems. The main goal of each problem is to fill in the last missing part of the matrix by solving which pattern or shape comes next. Participants need to solve as many matrices as possible in 10 minutes. This measure was specifically used as it provided a new digitized version of traditional non-verbal intelligence tests.

2) The Interaction Part Questionnaires: Here, we present the measures used in the Interaction part of the dataset. Over different stages, we collect different types of measures. Namely, following Principle **A1.P1**, we collected: 1) Expectation Measures, 2) Perception Measures (i.e., person, situation, and team perceptions), 3) Partner Selection Decisions Measures, and 4) Reflection Measures.

The Expectations measures were presented before the Conversation and Collaboration Stage, to capture people's expectations about the interaction they were about to have. Person Perception of Warmth and Competence was presented after the Impression, Conversation, and Collaboration Stages, while the remaining perception measures were repeated two times, once after the Conversation Stage and once after the Collaboration Stage. Partner Selection decisions were presented after the Impression and Conversation, and before the Collaboration Stage. See Fig. 1 to inspect the visualization and the list of measures used at every set of the Interaction part.

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

Expectation Measures (XM): In alignment with Principle **A1.P1**, we collected measures describing individuals' expectations w.r.t. the next upcoming interaction phase they were scheduled to engage in (i.e., either a Conversation or Collaboration Stage).

- *Expected Affect* was measured by answering the question *"How do you expect this conversation to feel?"* by using the Affect Button [61]. Affect Button is a validated measure for measuring affective states and judgments in terms of the Pleasure-Arousal-Dominance (PAD) model of affect, comprising continuous values for each of the three dimensions: *pleasure* (i.e., pleasure/discomfort), *arousal* (i.e., high/low level of bodily excitement), and *dominance* (i.e., high/low control over the situation). Ratings for each dimension are continuous in the interval [-1,1].
- *Expected Team perceptions* were measured by asking individuals what they expected from the interaction concerning Team Cohesion and Entitativity. We adjusted two questions from the Cohesion and Entitativity measures (Example: "*We will get along well with each other.*"; see the section on *Team Evaluation Measures* below for additional details).

Evaluation Measures (EM): In alignment with Principle **A1.P2**, we additionally collected measures describing how people perceived their partners and the situation, and how well they interacted.

- *Person perceptions of Warmth and Competence (W/C) and Personality* were measured as perceptions of other participants on warmth (i.e., trustworthiness and sociability) and competence, as measured in [62] and [63]. All three components were measured on three-item scales, respectively. *Person Perceptions of Personality* were measured with a six-item scale for measuring personality using the HEXACO model [64]. Each item was measured on a 9-point scale, representing an end of the spectrum for each personality trait from 1, *"not really lively and socially withdrawn, with low self-esteem"* to 9,*"lively, sociable, socially bold, with high social self-esteem"*.
- *Situation perceptions* were measured using the 10-item version of the Situational Interdependence Scale (SIS; [22]) and two one-item measures of Rapport and In syncness. The SIS was used to measure perceptions of situations on five dimensions of interdependence: mutual dependence, conflict of interest, information certainty, future interdependence, and relative power. On each item, participants needed to indicate their level of agreement (1 – "Strongly Disagree", 5 – "Strongly Agree"). Example of the items used for measuring Rapport: "*The interaction had good rapport (i.e., we understood each other and communicated well).*" and In-syncness *"We were in sync/in tune with each other."*). Each item was measured on a 5-point Likert scale (1 – *"Strongly Disagree"* to 5 – *"Strongly Agree"*). These items were hand-crafted following prior literature (see [65]).
- *Team perceptions* were measured through measures of cohesion and entitativity. Both measures focus on perceptions of a social group as a cohesive "unit". Each item was measured on a 5-point Likert scale (1 – *"Strongly Disagree"* to 5 – *"Strongly Agree"*). *Task and Social Cohesion* was measured using a modified two- or four-item measure used in [53]. Sample items included *"We got along well with each other"* (Social cohesion) and *"We both contributed to the mutual task"* (Task cohesion). Task cohesion was only measured after the Collaboration stage, as in the Conversation Stage, there was no task to report on. *Entitativity* was measured using a modified two-item measure used in [54]. Higher values indicated a higher level of cohesion.
- *Affect* was measured using the Affect Button [61] when answering the question *"How did this conversation make you feel?"*. As mentioned previously, AffectButton is a well-established measure used for measuring affect, and provides a more interactive interface for reporting experienced affect, compared to other measures for PAD (e.g., Self-Assessment Manikin [66]). Ratings for each dimension are continuous in the interval [-1,1].

Reflection Measures (RM) These measures are taken post-interaction to collect information describing the possible impact of remembered moments on individuals' perceptions of partners and interactions (see Principle **A1.P1**). In particular, here we asked participants to recall and reflect on the moment they remembered most clearly from the interaction they had just completed. We then provided them with the following measures:

- *Remembered Moment description* comprise a series of measures where individuals describe a) the content of the moment they remembered using free text (minimum 12 characters, up to 4 sentences), b) a reason for why they believe they remembered this specific moment (minimum 12 characters, up to 4 sentences), and c) the relative time at which they believe this moment took place using a Visual Analog Scale (slider with marks for "Beginning", "Middle", and "End"; mapping to an integer interval [0,100]).
- *Impact Reflections on Team Perceptions* asks participants to reflect on the impact they believe this moment had on elements of team perceptions related to Task and Social Cohesion (see Team Perception Measures above). Notably, these items comprised 5-point Likert scales that were anchored to focus on relative impact, not on absolute levels (e.g., "This moment resulted in us getting along... with each other", with viable options being "much worse", "a little worse", "neither better nor worse", "a little better", "much better").
- *Remembered Affect* is a measure asking participants to provide ratings in terms of Pleasure, Arousal, and Dominance with the AffectButton in response to the prompt of "How did experiencing this moment feel?". Ratings for each dimension are continuous in the interval [-1,1].

Partner Selection Measures

- *Partner Selection and Selection Strength* were measured with one item (i.e., "Would you want to do the Joint task/Feud task with this person?"). Note that we used different names for the Joint task in the experiment, calling it a Decision task. This was done so that we did not prime the participants with suggestive task names. This component was measured after the Impression stage, Conversation Stage, and Cooperation stage. On each item, participants

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

were asked to make a selection (0 – *"no"*; 1 – *"yes"*). For those who stated they wanted to choose a particular person, a follow-up question appeared asking them to rate the strength of the decision (i.e., *"How strongly do you want to be paired with this person?"*) (0 – *"Not at all"*; 100 – *"Very much"*). Additionally,

- *Willingness for future interaction* was measured with a one-item measure. The item prompted participants to decide if they wanted to meet with this person in real life after the study (0 – no; 1 – yes). Sample of the item used - *"Since you didn't have the chance to interact with this person for a longer period, you will have a chance to arrange another day to meet up with this person in real life and get to know them more. Would you be willing to meet up with this person in real life?"*). This item was used in addition to partner selection decisions to see whether people would also want to continue interacting with participants outside of the study.

Cooperation and Collaboration Measures

- *Cooperation* was measured in the Joint Tasks, where the MUs that were exchanged in the Joint Trust Task were used as a proxy of cooperative behavior. While there was no direct equivalent for this in the Joint Competence Task, participants' MU earnings served as a measure of task performance.
- *Collaboration* was primarily evaluated through task performance in the Feud Task, where both the number of correct responses and the sequence in which they were provided were considered.

3) Frontal Photographs: Except for questionnaires, we also collected frontal photographs of each participant, which were taken during the Impression Stage. These photographs were taken with their web cameras. Participants were instructed to remove any objects from their faces, such as glasses or visible accessories (e.g., earrings), and to make a neutral face. The photographs were collected to present them to other individuals before the conversations and reduce the effect of physical appearance in the conversation (i.e., we wanted participants to know with whom they would have the conversation).

4) Video Recordings: We collected audio-video recordings from a frontal view with a resolution of at least 640x480 and an audio sampling rate of 48,000 Hz. Two types of recordings were made—local and received—to account for the effects of latency and internet connection disruptions on audio-video quality. Local recordings captured the signal from the participant's own camera and microphone, representing the ideal signal before transmission, but could still be affected by latency and compression. Received recordings captured the signal of the participant's conversation partner as it arrived over the internet, subject to latency, compression, and other potential disruptions. For each participant, we aimed to collect one local recording and one received a recording from their conversation partner, resulting in approximately four audio-video recordings per interaction. During the conversation and collaboration sessions, participants saw the recording of their partner, while their own local recording was displayed in the corner of the screen. This setup allowed participants to focus on their conversation partner while remaining familiar with the technology, following common video conferencing software design patterns.

5) Web Application for Data Collection: For running the data collection procedure in the Interaction part, we opted for a solution combining a flow of survey elements created with the widely used software QUALTRICS, with a custom software solution for video conferencing and participant management. Our motivation for this choice was that it offered us a) flexibility in implementing our data collection protocol (e.g., by combining the substantial survey toolkit available in Qualtrics with custom management of participant interactions), while b) providing us with confidence in the privacy measures for data collection (e.g., by ensuring that no commercial third-party entity has access to data related to participants' video recordings).

Concretely, our overall distributed application integrates 1) a browser-based client structured around the Qualtrics Platform, together with 2) a custom server application handling management of participants and video recordings, and 3) relies on an instance of the JANUS WEBRTC SERVER[1] to facilitate video calls. Both of the latter servers were running under the full control of a research institution associated with the authors. Apart from data collection, the overall application fulfilled multiple functions: 1) Participant Monitoring (tracking participants' progress throughout the protocol), 2) Participant Management (e.g., control of the study flow and participant matching), 3) Communication with participants (e.g., through a bi-directional chat functionality with the authors and announcements; no chat between participants), and 4) Facilitating video calls, their recording locally in clients on both ends of the conversation, and automatic upload of to our servers (However, fallback procedures for participants to manually back up their recordings were in place, e.g., via an upload form).

Complexities related to the round-robin design made features related to *Participant Monitoring* and *Participant Management* especially vital requirements for our data collection application. For example, given the decentralized nature of participants' interactions with the survey client, it was challenging to ensure that they could successfully meet each other at specific moments in the protocol without progressing too fast or too slow. Similarly, remote participation in a diverse subject pool also bears the risks of potential technical issues that participants don't know how to take care of and may need to communicate with the experimenter about. Notably, the round-robin design is particularly vulnerable to delays since it requires each participant to interact with all other participants over different rounds. This interdependency means that any delay for one participant can have consequences for the remainder of the participants. Our application aimed to minimize such detrimental effects (e.g., by providing a forced synchronization mechanism, preventing progress for very fast participants until everyone in a group had cleared a certain stage).

Overall, this distributed application enabled us to present participants with an interface where all stages of the study are seamlessly integrated. Importantly, it guided them through the entire study without requiring a change between applications

[1]https://janus.conf.meetecho.com/

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

(e.g., interacting with a visibly different platform for video calls; in our setup, they were fully facilitated by the web browser being used). However, despite these benefits for participants, the integration of the different components proved not always very stable and complex to maintain and expand. Learning from these lessons, we have refined our prototype into an open tool for the community under the label OPENVIMO.

D. Ethics Statement

Due to the sensitivity of the data, multiple questions were asked to ensure that participants understood the terms and conditions they were agreeing to. Thus, a separate set of questions was used to see whether participants wanted to participate in the study and consent to the usage of their data in aggregated forms for publication purposes. Following these questions, another set of questions was used to provide consent for sharing the data in their raw form with other researchers. The data management plan, as well as the procedure for collecting data, and the consent form, were approved by the Ethical Committee at the Vrije University of Amsterdam (ID: VCWE-2021-168).

V. Dataset Curation

Originally, 297 participants completed both sessions of the study. In addition to only including participants who took part in both sessions, we created a curated version of the dataset.

Due to technical and study-related issues, such as participants withdrawing mid-study, corrupted videos, and missing evaluations, further filtering was required to produce the final clean version of the PARSEL dataset. Additionally, to make the dataset more user-friendly, we also performed some basic preprocessing steps, such as converting the format of video and audio recordings. These steps were taken to ensure the final version of the PARSEL dataset is accessible and ready for use. For transparency reasons, we document all of the steps in this section.

A. Data Filtering

Except for removing the participants who were not participating in both parts of the study, we filtered instances where photographs or audio-video recordings were corrupted (e.g., no visible image or audio-video recordings lasting for less than a minute). These technical problems can still occur during the remainder of the experiment. Therefore, we put a protocol in place to clean the recorded data. Additionally, we also removed the recordings with incomplete evaluation entries (e.g., if there was a conversation but the participant did not rate it).

1) Removing Evaluation Ratings With Corrupted Web Recordings and Incomplete Entries:

Removals from Impression Stage: Overall, the Impression Stage resulted in 1436 evaluation reports. This number was reduced to the final set of 1341 evaluation ratings after removing: a) incomplete evaluation ratings (n = 53), b) evaluation ratings of participant's pictures that did not provide any ratings of other participants, or evaluation ratings of a blank picture (n = 42), which occurred due to technical issues with the camera.

Removals from Conversation Stage: Every dyadic conversation consisted of two videos (one video representing one participant (P1) and the other video representing the other (P2)). Every video was matched with perceptions from the participants who were shown in the video. Overall, 1244 post-conversation perceptions were recorded. However, after removing audio-video recordings of conversations that were below one minute and conversations that had incomplete evaluation entries, this number was reduced to 1198 evaluations. The threshold of one minute was used, as it was thought that videos lasting less than one minute ended due to technical issues. Additionally, the one-minute threshold was used because it was believed that it provided individuals enough time to form an impression about the conversation and the situation.

Further, we removed any instances where participants did not have an evaluation for the same participant in the Impression and the Conversation stage, which decreased the number of evaluations to 1193, and removed any Conversation evaluations that did not have a corresponding decision in the Cooperation stage, decreasing the number of evaluations to 1149. This can occur either if a participant has dropped out before the Cooperation stage or if a scheduled conversation has not happened for certain dyads.

Lastly, we removed evaluations that were missing due to conversations not happening (n = 3) or evaluations where recorded conversations from both participants, either the Receiver or Local, were not available (n = 44). The unavailability of recordings from both participants prevented us from reconstructing the dyadic conversation. Thus, these evaluations were removed. This resulted in the final 1102 evaluations in the Conversation stage. For the full cleaning flow, see Appendix, Fig. 1.

Removals from Cooperation Stage: In total, 1310 responses were recorded for both Joint Tasks. However, when removing responses of pairings that were not available in the impression and conversation, this number decreased to 1102 responses.

Removals from Collaboration Stage: Overall, 1133 evaluations were recorded in the Collaboration stage. This set of evaluations was reduced to 930 evaluations when removing: a) videos below one minute and incomplete evaluation entries (n = 53), b) evaluations that were not shared between the previous and collaboration stages (n = 125) and c) removing evaluations where the collaborative interaction did not happen (n = 5) or the recordings of the collaborative interaction was not available for both participants (n = 20).

Due to technical issues, the stages had different numbers of observations, with the Collaboration stage having the least number of observations. Given the multi-stage design, each stage can be explored independently (e.g., looking only at the Conversation stage). However, when cleaning each stage, we enforced a cleaning criterion that all of the evaluations for all dyads needed to be shared among all of the stages. When interpreting data or modeling social phenomena of interest, these cleaning criteria acknowledge the context of the study (e.g., the participant's journey that happened before the Collaboration Stage). Specifically, we kept evaluations that had comparable study journeys (i.e., participants evaluated that person over all stages).

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

B. Data Processing

While creating the curated and cleaned version of the dataset, the following transformation of the original data was applied to the Questionnaire measures from the Intake and Interaction Part.

The Intake Part Measures Transformations: All responses on the evaluation questionnaires were aggregated by calculating average values for each corresponding questionnaire. For calculating the average values of all the Intake measures (i.e., social value orientation, personality traits, social anxiety, intelligence, and psychopathy), we used the corresponding published keys and documentation (see [27], [28], [55], [60]). The published keys provided us with instructions on how to properly calculate the averages for each scale. Additionally, they give information on whether any reverse-coded items need to be properly scored.

The Interaction Part Measures Transformations: All responses on the evaluation questionnaires were aggregated by calculating average values for each corresponding questionnaire. For calculating the average values of perceptions of warmth, competence, situational interdependence, social cohesion, and entitativity, the corresponding published keys and instructions were used (see [22], [53], [54], [61], [62], [63]).

1) Webcam Recordings: Additional processing steps were used on the webcam recordings to extract the (non)-verbal features to extend the data type provided in the final and curated dataset.

Transcoding Webcam Recordings: All audio-video recordings were transcoded from the original format (.webm) to a format that was supported by most Python packages dealing with the analysis of multi-modal input (.mp4). All raw recordings were transcoded to a predefined resolution format of 640*480, and a frame rate of 30 frames per second. We also extracted audio from all webcam recordings in the. wav format.

Extracting non-verbal behavioral features: To capture information about the facial and acoustic non-verbal behavioral features, two pre-trained models, OpenFace2.0 [67] and OpenSmile [68], were applied on the video and audio part of the webcam recordings. OpenFace automatically measures facial muscle movements as *action units (AU)* that correspond to specific facial expressions and is based on the Facial Action Coding Scheme (FACS). Specifically, OpenFace provides intensity measurements of 17 AU for each frame, ranging from 0 to 5 (where 0 represents no activation). These AU detect actions such as Inner Brow Raiser, Brow Lowerer, Upper Lid Raiser, Cheek Raiser, Nose Wrinkler, Lip Corner Depressor, Blink, and many more. Additionally, OpenFace provides measurements of head movement and eye gaze. Similarly, OpenSmile automatically measures the non-verbal characteristics of the speech. Here, we utilized a pre-defined set of features - ComParE2016, which includes 65 low-level audio descriptors [69], including pitch, intonation, jitter, shimmer, and many more.

To support a range of research goals, we included all extracted AUs from OpenFace, enabling researchers to select features most relevant to their hypotheses. However, when selecting the features, we acknowledge two caveats: (1) confidence with which OpenFace identified faces varies; However, estimates provided by the software are overall relatively high (M = 0.96, SD = 0.03, Min/Max = 0.70/0.96. Additionally, (2) lower-face AUs may be activated by speech, which could confound interpretations of their meaning as expressive behavior. We encourage researchers to make use of the confidence values provided when selecting their AU of choice and, if applicable, to consider integrating information from the raw data in terms of speech activity when interpreting movements in the lower face. Finally, we encourage researchers to consider the error distribution in time and concerning person-specific variations that are required for their research question. Carrying out their validation tests on a subset of the data in settings that vary significantly from the setups described in this paper is highly recommended.

VI. Descriptive Statistics of the Questionnaires

Understanding the qualities of the data set helps researchers assess its usability and identify potential opportunities. For instance, having a highly skewed distribution of certain participant characteristics might make the dataset less suitable for measuring its effect, due to constrained variance. On the other hand, being aware of the dataset characteristics can also help researchers be mindful when modeling the data (see the idea behind Dataset Cards [70]). Thus, this section aims to provide a concise overview of data patterns and distributions to better highlight the dataset's and participants' characteristics.

A. Intake Part Measures

All scales used in the Intake Part had good internal validity as measured by Cronbach's alpha. Including, Personality: α_{HH} =. 77; α_A =. 84; α_{EX} =. 84; α_E =. 78; α_C =. 80; α_O =. 79, Anxiety: α_{SIAS} =. 85, and Psychopathy: α_{PPTS} =. 86.

Demographics: In total, the curated version of the dataset included 297 participants (166 females; $M_{age} = 37.39$, $SD_{age} = 11.42$). The majority of the sample reported nationalities of the United Kingdom ($n = 227$), Italy ($n = 6$), Scotland ($n = 6$), India ($n = 5$), and Nigeria ($n = 5$). Other nationalities being reported were: Spain, Poland, Romania, Germany, Ireland, Portugal, Estonia, Australia, Bangladesh, United States, Lithuania, Malta, Malaysia, France, Namibia, Serbia, China, Russia, Ghana, and Ukraine. This indicates that our dataset captured different nationalities, while all of them lived in the U.K. at the time of the data collection. Participants were nested across 55 batches in the interaction part of the dataset.

Personality, Social Anxiety, Psychopathy, and Trust: The participants in the dataset exhibited variation across all six HEXACO personality traits, with scores spanning the full range of the 5-point Likert scale. Descriptive statistics (see Table II) demonstrate that individuals differed across each trait, reflecting a diverse spectrum of personality profiles within the sample. Similarly, on average, participants reported relatively low levels of social anxiety (SIAS; M = 2.40, SD = 0.90) and psychopathy (PPTS; $M = 2.45$, $SD = 0.54$). In contrast, mean scores on trust-related measures were high, particularly for benevolence (M = 4.04, SD = 0.68) and integrity (M = 4.22, SD = 0.55). All constructs were assessed using a 5-point Likert scale.

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

TABLE II
DESCRIPTIVE STATISTIC OF THE INTAKE QUESTIONNAIRES AND EVALUATION MEASURES IN THE INTERACTION PART

Variable	Measure	*Md*	*M*	*SD*	*Min/Max*
Intake Questionnaires					
Personality	Honesty-Humility	3.50	3.43	0.71	1.60/4.80
	Agreeableness	3.50	3.29	0.72	1.10/5.00
	Extraversion	3.20	3.23	0.76	1.30/4.80
	Conscientiousness	3.80	3.69	0.66	1.90/5.00
	Openess	3.70	3.69	0.71	1.50/5.00
	Emotionality	3.30	3.28	0.70	1.20/4.80
Social Anxiety	Social Anxiety	2.33	2.40	0.91	1.00/4.67
Psychopathy	Psychopathy	2.40	2.46	0.55	1.25/4.10
Trust	Benevolence	4.20	4.04	0.68	2.00/5.00
	Ability	3.85	3.82	0.63	1.50/5.00
	Propensity	3.25	3.27	0.46	1.50/4.63
	Integrity	4.33	4.22	0.55	2.17/5.00
Non-verbal Intelligence	UCMRT	11.00	10.61	5.11	0.00/22.00
Interaction Questionnaires					
Photograph Impression Stage					
Person Perceptions	Warmth	5.33	5.10	1.04	1.00/7.00
	Sociability	5.33	5.11	1.14	1.00/7.00
	Morality	5.33	5.09	1.09	1.00/7.00
	Dominance	4.00	\	\	1.00/7.00
	Similarity	4.00	\	\	1.00/7.00
	Attractiveness	5.00	\	\	1.00/7.00
	Likeability	5.00	\	\	1.00/7.00
	Competence	5.33	5.25	1.06	1.00/7.00
Partner Selection	Selection Strength	60.00	61.46	17.05	5.00/100.00
Conversation Stage					
Person Perceptions (W/C)	Warmth	6.17	6.01	0.84	1.00/7.00
	Sociability	6.33	6.09	0.90	1.00/7.00
	Morality	6.00	5.93	0.91	1.00/7.00
	Dominance	4.00	\	\	1.00/7.00
	Similarity	5.00	\	\	1.00/7.00
	Attractiveness	5.00	\	\	1.00/7.00
	Likeability	6.00	\	\	1.00/7.00
	Competence	6.00	5.84	0.93	1.00/7.00
Person Perceptions (Personality)	Honesty-Humility	7.00	\	\	1.00/9.00
	Emotionality	5.00	\	\	1.00/9.00
	Extraversion	7.00	\	\	1.00/9.00
	Agreeableness	7.00	\	\	1.00/9.00
	Conscientiousness	6.00	\	\	1.00/9.00
	Openness	6.00	\	\	1.00/9.00
Situation Perceptions	Mutual Dependence	4.00	3.83	0.79	1.00/5.00
	Conflict of Interest	2.00	2.00	0.76	1.00/5.00
	Future Interdependence	3.50	3.41	0.81	1.00/5.00
	Information Certainty	3.50	3.34	0.96	1.00/5.00
	Relative Power	3.00	3.10	0.65	1.00/5.00
Team Perceptions	Social Cohesion	4.50	4.26	0.66	1.00/5.00
Conversation Perceptions	In syncness	4.00	\	\	1.00/5.00
	Rapport	4.00	\	\	1.00/5.00
Partner Selection	Selection Strength	72.00	70.76	18.60	7.00/100.00
Collaboration Stage					
Person Perceptions (W/C)	Warmth	6.00	6.02	0.81	1.00/7.00
	Sociability	6.00	6.05	0.87	1.00/7.00
	Morality	6.00	5.99	0.86	1.00/7.00
	Dominance	4.00	\	\	1.00/7.00
	Similarity	5.00	\	\	1.00/7.00
	Attractiveness	5.00	\	\	1.00/7.00
	Likeability	6.00	\	\	1.00/7.00
	Competence	6.00	5.95	0.81	1.00/7.00
Person Perceptions (Personality)	Honesty-Humility	7.00	\	\	1.00/9.00
	Emotionality	4.00	\	\	1.00/9.00
	Extraversion	7.00	\	\	1.00/9.00
	Agreeableness	7.00	\	\	1.00/9.00
	Conscientiousness	7.00	\	\	1.00/9.00
	Openness	7.00	\	\	1.00/9.00
Situation Perceptions	Mutual Dependence	4.00	3.88	0.75	1.00/5.00
	Conflict of Interest	1.50	1.82	0.75	1.00/5.00
	Future Interdependence	3.00	3.40	0.82	1.00/5.00
	Information Certainty	4.00	3.81	0.86	1.00/5.00
	Relative Power	3.00	3.08	0.58	1.00/5.00
Team Perceptions	Social Cohesion	4.50	4.33	0.61	1.00/5.00
	Task Cohesion	5.00	4.53	0.59	1.00/5.00
	Entitativity	4.00	4.22	0.70	1.00/5.01
Conversation Perceptions	In syncness	4.00	\	\	1.00/5.00
	Rapport	4.00	\	\	1.00/5.00
Partner Selection	Selection Strength	73.00	71.25	18.94	1.00/100.00

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

Social Value Orientation was calculated using the formulas provided in [28]. The majority of the sample belongs to the Altruistic category (n = 181), followed by the Prosociality category (n = 75), the Individualism category (n = 36), and the Competitiveness category (n = 5). This indicates that most participants included in PARSEL have altruistic or prosocial tendencies towards other individuals.

Non-Verbal Intelligence was calculated as the sum of scores obtained in the UCMRT task. On average, participants solved 11 ($SD = 5.11$) tasks correctly, with 0.00 being the lowest score obtained and 22 the highest score obtained.

B. The Interaction Part Measures

1) Expectation Measures (XM):

Expected Affect in the conversation was expected to be relatively pleasurable ($M = 0.39$, $SD = 0.39$, $Md = .36$), and scored positive for dominance ($M = 0.34$, $SD = 0.49$, $Md = .40$), and negative on arousal ($M = -0.16$, $SD = 0.75$, $Md = -.29$). Similarly, before collaboration participants' affect expected the task to be relatively pleasurable ($M = 0.54$, $SD = 0.38$, $Md = 0.62$), positive on dominance ($M = 0.47$, $SD = 0.48$, $Md = 0.58$), and positive on arousal ($M = 0.11$, $SD = 0.73$, $Md = 0.28$).

Anticipated Team Perceptions were relatively high before conversations. On average, participants expected to get along well with their future conversation partners ($M = 3.99$, $SD = 0.62$, and $Md = 4.00$). Before the collaboration task, the anticipation of the social cohesion was also relatively high ($M = 4.38$, $SD = 0.61$, and $Md = 4.50$). This indicates that participants were already expecting to get along well with other participants before the interactions. This insight is important to consider, as expectations have been shown to impact one's behavior [71].

2) Evaluation Measures (EM): All measures showed high internal validity, including Person Perceptions of Warmth and Competence: $\alpha_{Warmth} = .91$; $\alpha_{Competence} = .91$), Situation perceptions: $\alpha_{MutualDependence} = .48$; $\alpha_{ConflictofInterest} = .62$; $\alpha_{FutureInterdependence} = .67$; $\alpha_{InformationCertainty} = .72$; $\alpha_{RelativePower} = .66$), and Team perceptions: $\alpha_{SocialCohesion} = .49$; $\alpha_{TaskCohesion} = .74$; $\alpha_{Entitativity} = .89$.

Person Perceptions of Warmth and Competence (W/C) and Personality were positive across all stages of the dataset. In the photograph impression stage (n = 1341) all participants rated their conversation partners positively on all dimensions of person perceptions, including warmth (sociability ($M = 5.11$, $SD = 1.41$) and morality ($M = 5.10$, $SD = 1.09$)), competence ($M = 5.25$, $SD = 1.06$), and attractiveness ($Md = 5.00$), but less so for dominance ($Md = 4.00$) and similarity ($Md = 4.00$). However, these perceptions were even more positive after the conversation and the collaboration task (see Table II for changes in perceptions). Median values of Person's Perceptions of Personality after the Conversation stage indicate that all personality traits were rated on the positive side of the distribution, including Honesty-Humility ($Md = 7.00$), Emotionality ($Md = 5.00$), Extraversion ($Md = 7.00$), Agreeableness ($Md = 7.00$), Conscientiousness ($Md = 6.00$), and Openness ($Md = 6.00$). Person Perceptions of Personality were measured on a 9-point scale. Similar values were observed in the Collaboration Stage (see Table II). These results indicate that although participants' perceptions were, on average, on the positive side, their perceptions strengthened in positivity after the initial conversation and stayed similar after the Collaboration Stage.

Situation Perceptions indicate that interactions during the Conversation Stage were evaluated to be low on conflict of interest ($M = 2.00$, $SD = 0.76$), both speakers had perceptions that they had equal power in the situation ($M = 3.10$, $SD = 0.65$), relatively high on mutual dependence (i.e., actions of each participant affected the actions of the other in the situation) ($M = 3.83$, $SD = 0.79$), and moderate on future interdependence ($M = 3.41$, $SD = 0.81$) and information certainty (i.e., knowing what the other person in the conversation wants and needs) ($M = 3.34$, $SD = 0.86$). These values stayed relatively similar to the ones collected after the Collaborative Stage (see Table II). Except for these, participants also reported having relatively high rapport ($Md = 4.00$) and in-syncness ($Md = 4.00$) after the Conversation Stage, which remained stable after the Collaboration Stage.

Descriptive statistics of the situation perceptions are in line with our assumption that interactions with strangers would usually be perceived to be equal in power, more friendly, and mutually dependent, as these are the characteristics that are common for informal and casual conversations. Additionally, we did not expect participants to have high information certainty, as they interacted with strangers.

Team Perceptions measured after the Conversation Stage indicated that participants reported getting along well with each other (Social Cohesion: $M = 4.26$, $SD = 0.66$). These results were also similar to evaluations of team perceptions after the Collaboration stage (see Table II). Given the fact that these results were on the higher end of the distribution, ranging from one to five, we can conclude that participants had a positive perception of cohesiveness.

3) Partner Selection Measures: Results showed that when selecting their partners based on observing a photograph of them, 70.25 % (n = 942) of 1341 evaluations indicated that individuals wanted to be matched with that participant for a specific cooperative task. Similar values were observed after the conversation, where out of 1102 observations, 71.51 % (n = 788) of the participants wanted to be matched with the participant for a specific cooperative task. Additionally, in half of the observations, participants were willing to meet their conversation partners outside the study (n = 537). Lastly, after the cooperative task, we asked participants to select a partner for a collaborative task. Out of 933 observations in 690 of the perceptions, participants wanted to be matched with another participant for the collaboration task. Moreover, in half of the observations, participants were willing to meet their conversation partners outside the study (n = 413). These observations are relevant as they indicate that participants were selective in their choices, but they were not too rigorous in the selection, as they wanted to interact with more than one participant.

Reflection moments are not included in the current analysis, due to their qualitative nature and the restricted space of the paper format.

4) Cooperative and Collaborative Behavior: In the Joint Trust Task, participants shared on average 6.28 MU from 10 MU in total ($SD = 3.55$ MU). This average was equivalent to 3.13

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

pounds. The values participants shared with others ranged from 0 MU to 10 MU, covering all the possible values. This indicates a high variability in cooperative behaviors in the Joint Trust Task. However, the average value indicates that participants shared half of their endowment with other participants, but the standard deviation indicates that there were differences.

This observation was somewhat expected, as participants interacted with each other before the task, increasing cooperative behavior, which was already observed in prior literature [72].

In the collaborative task, 1.18 % (out of 930) of the observations had all correct answers. The majority of interactions had 2/3 out of 6 answers correct (n = 284) or 1/2 answers correct (n = 255). The remaining participants had 5/6 answers correct (n = 109), and 1/3 out of 6 answers correct (n = 169). 49 participants did not have any correct answers.

The distribution of results across different performance levels (from no correct answers to nearly perfect scores) suggests that the task effectively differentiated participants' abilities. It could also reflect varying degrees of collaboration effectiveness, which can be used for further exploration.

C. Statistics of Recorded Behavior

1) Facial Expressions: For each video recorded in the Conversation (n = 2101 videos) and Collaboration Stage (n = 1943 videos), we calculated the mean and standard deviation of the intensities of all AU that are extracted using OpenFace. The average values of AU intensities in the conversation session ranged from 0.49 to 1.65 out of a maximal intensity of 5. Most importantly, all AUs were, on average, captured with high confidence ($M = 0.96$; $SD = 0.03$), indicating a reliable annotation technique.

Similarly, the intensities of all AU in the collaboration session ranged from average intensity between 0.43 and 1.42, which were captured with a high overall confidence ($M = 0.89$; $SD = 0.16$). The average intensities were calculated only over occurrences where the AU in question was detected.

VII. Examples of Potential Uses of PARSEL

The primary goal of data collection for PARSEL was to address literature gaps and explore how individuals select partners for cooperative and collaborative tasks, including the role of person perception in partner selection and modeling these processes, as discussed in the next section. It is worth noting that, while these examples demonstrate tasks for which PARSEL was already used, only certain models, pipelines, and types of train-test splitting were used. Thus, besides showcasing the usability, these examples also highlight opportunities for other researchers to build on them, improving and optimizing the machine learning pipelines, leading to better model performance.

A. Predicting Partner Selection From Person Perceptions of Warmth and Competence

Using PARSEL, the authors' prior work has shown that there is a significant relationship and predictive value of a person's perceptions of warmth and competence when predicting the likelihood of being selected [73]. Specifically, participants perceived to be higher on warmth and competence were also more likely to be selected as partners. Additionally, this relationship was moderated by task requirements, where perceptions of competence were more predictive of partner selection outcomes in the Joint Competence Task. Conversely, for the Joint Trust Task, warmth was important. Interestingly, our results show that perceptions of warmth were predictive of partner selection in both tasks. These findings align with previous literature that established people use person perceptions when planning future actions [14] and take task requirements into account when selecting partners [21].

B. Modeling Person Perceptions of Warmth and Competence

Using PARSEL, in the same paper [73], we also explored the feasibility of using machine learning to predict person perceptions, specifically warmth and competence, based on nonverbal behavioral features extracted from introductory conversations. This modeling effort was motivated by the importance of warmth and competence in social judgments [74].

We developed models using facial (OpenFace AUs) and vocal (OpenSmile) features as input and found that nonverbal cues offered modest, but significant predictive value for warmth perceptions over the dummy baseline, while predictions of competence were notably weaker and not significantly better. These findings are in line with prior research suggesting that naturally occurring nonverbal behavior may have limited predictive power when it comes to certain social perceptions, such as Extraversion or competence [75].

Despite these promising initial findings, more research is needed to explore how other cues could be used to improve the model's performance, such as the linguistic content of the speech or conversation dynamics (e.g., the temporal and structural pattern of the conversational exchange). Further research could also use PARSEL to directly build on our findings in Matej Hrkalovic et al. (2025) [73] by exploring the benefits of different machine-learning pipelines for our investigated task formulations.

C. Modeling Partner Selection for Cooperative Tasks

In certain situations, gathering people's perceptions of warmth and competence may prove challenging. Therefore, another interesting possibility is to investigate whether we can predict partner selection directly through non-verbal behaviors. Thus, this example aimed to explore whether partner selection can be modeled only based on non-verbal behaviors.

In this study, we utilized facial and acoustic features as inputs and post-conversation partner selection decisions as outputs. Since the setup involved dyadic interactions, we limited our analysis to cases where participants were classified as either choosers or those being chosen. Our focus here was specifically on whether the behaviors of the participants being chosen could predict the partner selection decisions made by the choosers. This resulted in a reduced dataset of 551 observations. This distinction was made to enable future research where researchers can extend these models with coordination measures between

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

both participants, where role differentiation is essential. We then applied three classification models: a baseline majority classifier, a RIDGE classifier, and an SVC classifier. The task was a binary classification.

1) Machine Learning Pipeline: The machine learning pipeline included all of the pre-processing stages (see Data Filtering & Data Curation). After extracting facial and acoustic cues, all feature vectors were additionally zero-padded to the maximum vector length (5400 $\times$ 1), which was equivalent to a 3-minute duration of audio-video recordings. Each recording was represented as a matrix of 5400 rows (timepoints) x 114 columns (OpenSmile (columns = 65) and OpenFace features (columns = 49). Then, the matrix of a vector of features (e.g., time series) was concatenated together and transformed using the MiniROCKET algorithm. MiniRocket uses random convolutional kernels to transform the time series data. This transformation captures essential patterns and features, which are then used for classification [76]. After the transformation, each observation (i.e., audio-video recording) was represented as a vector of 9996 time-series features. Lastly, the matrix (9996 features x 551 observations) was fed into the corresponding classifier. This pipeline was already used in previous studies and was successful for modeling perceptions of interdependence [77] and for modeling perceptions of warmth and competence [73].

2) Model Evaluation: To estimate the model's performance and select the best corresponding model hyperparameters, a 5-fold nested cross-validation procedure was used. Due to the nested nature of the experimental setup, each data fold was determined by ensuring that data within the same group were placed in the same fold. Due to the imbalanced classes, *Balanced Accuracy* was used as a performance metric. It is computed as the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate).

3) Results: All models were statistically compared against the performance score of a majority classifier (i.e., Balanced Accuracy =. 50). For all comparisons, we used a t-test. Comparisons showed that while the performance of the RIDGE classifier was higher compared to the majority classifier (Balanced Accuracy =. 53; $SD = 0.02$), this difference was not statistically significant ($t(4) = -2.38$, p =. 076. On the other hand, the performance of the SVC model (Balanced Accuracy = .56; $SD = 0.01$) was statistically better when compared with the majority classifier ($t(4) = -11.34$, $p <. 001$).

VIII. Potential Uses in Future Research

Except for modeling partner selection or person perceptions from non-verbal behaviors, PARSEL can be used for addressing other research questions and modeling tasks. Here, we describe some additional research endeavors for which we believe PARSEL can be useful. Concretely, we outline a series of scenarios (tasks), providing for each of them 1) a clear *Task Description* and contextualization, 2) an argument of *Task Relevance* – outlining why this task can be considered a meaningful and impactful research enterprise –, and 3) a summary of the *PARSEL resources* that can support associated research.

A. Predictive Modeling of Collaborative Performance From Prior Behavior

Task Description: People's behavior in previous interactions affects how they are perceived by others and can influence trust and, ultimately, the group's overall performance [78]. While most research has centered on developing models that predict collaborative performance by analyzing behaviors from these same interactions, less attention has been given to investigating whether systems can benefit from analyzing behavior from past interactions to predict future collaborative performance.

Task Relevance: Developing such models could help pair individuals more effectively for future collaborations. By using data from past interactions, we could recommend partners who are more likely to perform well together. Detecting potential success or failure early in a collaboration could also provide real-time feedback, enabling teams to adjust or halt interactions that are unlikely to succeed.

PARSEL resources: PARSEL contains both audio and video recordings from initial conversations and later stages of collaboration, along with objective measures of collaboration performance. It also includes first-person evaluations of each participant after their initial interactions, which can be additionally used to train more comprehensive models.

B. Modeling Expected Affective and Conversational Experience

Task Description: Socially Intelligent Systems are envisioned to support humans in conversational settings, e.g., by fostering effective collaboration [79]. In service of these goals, systems try to analyze the behavior that people display during conversation to estimate aspects of conversational experience, either as evaluated by participants at the moment throughout the interaction or in reflection after the fact (see Dudzik& Broekens for an in-depth discussion in the context of emotional self-report [80]). However, one aspect that has not been considered extensively in this line of research is inferring people's *expectations* of affective and cohesive experience before a conversation from the behavior they display.

Task Relevance: Expectations of affect and conversational experience play a crucial role in shaping human social behavior and decision-making [81], [82]. As such, inferring not only how a conversation might be related to human experience but also the extent to which this experience relates to prior expectations might prove invaluable for offering intelligent strategies for social support. For example, fostering effective collaboration with people who share negative expectations from the event onset might require different interventions for a positive outcome compared to a scenario where something went awry during the conversation itself. Additionally, exposing humans to information about their anticipatory emotions and the role they could have played in their experience, can offer a valuable element for fostering reflection and learning (see Chen et al. [83] for a relevant example), as well as a means in service of trustworthy AI through explainability [84].

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

PARSEL resources: The dataset contains rich data on conversational behavior and measures of expected affect and cohesiveness (Expectation Measures). Together, these can serve as crucial data in support of this type of modeling research. Moreover, it captures variation in terms of different settings that could be explored (e.g., the different *Joint Task* conditions or phases of partner selection)

IX. Limitations

Despite its strengths, PARSEL also has some limitations. Here, we shortly outline several of them.

Ecological Validity: One limitation is that participants rated qualities of the person, conversation, and situation before and after each interaction, which may reduce ecological validity since such explicit ratings aren't typical in real life. We assumed a linear order in how evaluation processes play out —expectations first, then interaction evaluations, and finally recollections. While this ordering could reflect natural processes, these judgments may be more implicitly integrated in decision-making. To mitigate this, to some extent, our survey items included multiple characteristics to minimize the influence of any single quality on behavior, such as partner preference.

Constrained Cooperative Settings: We used only two specific cooperative tasks related to warmth and competence traits in the dataset. Therefore, the dataset we collected is specific to the settings of these two tasks. However, we used validated tasks that have been used in previous literature to examine cooperative behavior [49], [50]. Thus, we believe that the cooperative settings are representative of the task at hand. However, more research is needed to extend the variety of tasks that can be used to measure cooperative behavior and, therefore, consider the role of situational context more broadly on partner selection and perception.

X. Conclusion

In this paper, we introduced PARSEL, a novel and richly annotated multimodal dataset designed to advance research on human partner selection in cooperative and collaborative settings. While prior work has underscored the importance of selecting suitable partners, little attention has been given to how the sequential, perception-driven decision-making process individuals undergo during such selection. PARSEL fills this gap by capturing the complexity of this process through a combination of audiovisual recordings, validated self-report measures, and behavioral outcomes across various interaction stages.

Beyond modeling perceptions such as warmth and competence, the dataset provides a foundation for exploring a wide range of socio-cognitive and behavioral phenomena, including interaction dynamics, non-verbal synchrony, situation and team perceptions, and their links to tangible collaboration outcomes. By validating the dataset through expected patterns derived from existing social psychology literature and demonstrating its application in modeling partner selection, we establish its relevance and utility for both computational and social science research.

We envision PARSEL as a resource that will not only facilitate the development of intelligent systems capable of supporting human decision-making in social contexts but also serve as a valuable tool for deepening our theoretical understanding of human collaboration. Ultimately, this work aims to bridge the gap between empirical psychological insights and practical computational solutions, paving the way for more mindful, effective, and supported human interactions.

Dataset Availability

The dataset is hosted by Delft University of Technology and the Human-Oriented Machine Intelligence Unit and is available upon request and after signing an End User License Agreement. To indicate your interest in the data, please use this form. Researchers can choose to access 'raw media' (that is, raw video and audio files for each speaker), "processed media" (i.e., processed feature files for each conversation), and/or "no media" (i.e., just the survey data). After filling in the forms, an End-User License Agreement will be sent, which needs to be signed. After signing, the researchers get a link and a password to download the data. You can contact: parsel_ewi@tudelft.nl for all inquiries about the dataset.

References

[1] B. C. Feeney and N. L. Collins, "A new look at social support: A theoretical perspective on thriving through relationships," *Pers. Social Psychol. Rev.*, vol. 19, no. 2, pp. 113–147, 2015.
[2] R. Noë and P. Hammerstein, "Biological markets: Supply and demand determine the effect of partner choice in cooperation, mutualism and mating," *Behav. Ecol. Sociobiol.*, vol. 35, pp. 1–11, 1994.
[3] P. Barclay, "Biological markets and the effects of partner choice on cooperation and friendship," *Curr. Opin. Psychol.*, vol. 7, pp. 33–38, 2016.
[4] J. Martin, L. Young, and K. McAuliffe, "The psychology of partner choice," 2019. [Online]. Available: https://doi.org/10.31234/osf.io/weqhz
[5] S. Siuda, T. Schlösser, and D. Fetchenhauer, "Do we know whom to trust? A review on trustworthiness detection accuracy," *Int. Rev. Social Psychol.*, vol. 35, 2022, Art. no. 20.
[6] B. Jaeger, B. Oud, T. Williams, E. G. Krumhuber, E. Fehr, and J. B. Engelmann, "Can people detect the trustworthiness of strangers based on their facial appearance?," *Evol. Hum. Behav.*, vol. 43, no. 4, pp. 296–303, 2022.
[7] Y. Bian, C. Zhou, Y. Zhang, J. Liu, J. Sheng, and Y.-J. Liu, "Focus on cooperation: A face-to-face VR serious game for relationship enhancement," *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 913–928, Jul.–Sep. 2024.
[8] A. B. Eisenbruch and M. M. Krasnow, "Why warmth matters more than competence: A new evolutionary approach," *Perspectives Psychol. Sci.*, vol. 17, no. 6, pp. 1604–1623, 2022.
[9] T. M. Hrkalovic, "Designing hybrid intelligence techniques for facilitating collaboration informed by social science," in *Proc. 2022 Int. Conf. Multimodal Interaction*, 2022, pp. 679–684.
[10] E. Weisz, D. C. Ong, R. W. Carlson, and J. Zaki, "Building empathy through motivation-based interventions," *Emotion*, vol. 21, no. 5, 2021, Art. no. 990.
[11] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Trans. Multimedia*, vol. 16, pp. 1075–1089, 2014.
[12] L. Cabrera-Quiros, E. Gedik, and H. Hung, "Multimodal self-assessed personality estimation during crowded mingle scenarios using wearables devices and cameras," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 46–59, Jan.–Mar. 2022.
[13] H. J. Escalante et al., "Modeling, recognizing, and explaining apparent personality from videos," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 894–911, Apr.–Jun. 2022.
[14] M. L. Patterson, "Invited article: A parallel process model of nonverbal communication," *J. Nonverbal Behav.*, vol. 19, pp. 3–29, 1995.

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

[15] O. FeldmanHall and A. Shenhav, "Resolving uncertainty in a social world," *Nature Hum. Behav.*, vol. 3, no. 5, pp. 426–435, 2019.
[16] K. R. Scherer, "Appraisal considered as a process of multilevel sequential checking," in *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford, U.K.: Oxford Univ. Press, 2001, pp. 92–120.
[17] D. A. Redelmeier and D. Kahneman, "Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures," *Pain*, vol. 66, no. 1, pp. 3–8, 1996.
[18] D. Wirtz, J. Kruger, C. N. Scollon, and E. Diener, "What to do on spring break?," *Psychol. Sci.*, vol. 14, no. 5, pp. 520–524, Sep. 2003, doi: 10.1111/1467-9280.03455.
[19] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*. New York, NY, USA: Cambridge Univ. Press, 1990.
[20] P. Barclay, "Strategies for cooperation in biological markets, especially for humans," *Evol. Hum. Behav.*, vol. 34, no. 3, pp. 164–175, 2013.
[21] J. L. Clark, M. C. Green, and J. J. Simons, "Narrative warmth and quantitative competence: Message type affects impressions of a speaker," *PLoS One*, vol. 14, no. 12, 2019, Art. no. e0226713.
[22] F. H. Gerpott, D. Balliet, S. Columbus, C. Molho, and R. E. d. Vries, "How do people think about interdependence? A multidimensional model of subjective outcome interdependence," *J. Pers. Social Psychol.*, vol. 115, no. 4, 2018, Art. no. 716.
[23] E. Baggs, "All affordances are social: Foundations of a Gibsonian social ontology," *Ecological Psychol.*, vol. 33, no. 3/4, pp. 257–278, 2021.
[24] B. Dudzik, H. Hung, M. Neerincx, and J. Broekens, "Collecting mementos: A multimodal dataset for context-sensitive modeling of affect and memory processing in responses to videos," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1249–1266, Apr.–Jun. 2023.
[25] A. Augustine and S. Hemenover, "Extraversion, social interaction, and affect repair," in *Encyclopedia of the Sciences of Learning*. Boston, MA, USA: Springer, 2012, pp. 1253–1255.
[26] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Acad. Manage. Rev.*, vol. 20, no. 3, pp. 709–734, 1995.
[27] T. Wu, Y. Luo, L. S. Broster, R. Gu, and Y.-j. Luo, "The impact of anxiety on social decision-making: Behavioral and electrodermal findings," *Social Neurosci.*, vol. 8, no. 1, pp. 11–21, 2013.
[28] R. O. Murphy, K. A. Ackermann, and M. J. Handgraaf, "Measuring social value orientation," *Judgment Decis. Mak.*, vol. 6, no. 8, pp. 771–781, 2011.
[29] S. E. Tait and D. Jeske, "Hello stranger!: Trust and self-disclosure effects on online information sharing," *Int. J. Cyber Behav., Psychol. Learn.*, vol. 5, no. 1, pp. 42–55, 2015.
[30] J. Ernst et al., "Social interactions in everyday life of socially anxious adolescents: Effects on mental state, anxiety, and depression," *Res. Child Adolesc. Psychopathol.*, vol. 52, no. 2, pp. 207–222, 2024.
[31] M. Calder et al., "Computational modelling for decision-making: Where, why, what, who and how," *Roy. Soc. Open Sci.*, vol. 5, no. 6, 2018, Art. no. 172096.
[32] Z. Akata et al., "A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence," *Computer*, vol. 53, no. 08, pp. 18–28, 2020.
[33] C. Oertel, K. A. F. Mora, S. Sheikhi, J.-M. Odobez, and J. Gustafson, "Who will get the grant? A multimodal corpus for the analysis of conversational behaviours in group interviews," in *Proc. 2014 Workshop Understanding Model. Multiparty, Multimodal Interact.*, 2014, pp. 27–32.
[34] A. Janin et al., "The ICSI meeting corpus," in *Proc. 2003 IEEE Int. Conf. Acoust., Speech, Signal Process., 2003 Proc.*, 2003, pp. I–I.
[35] W. Kraaij, T. Hain, M. Lincoln, and W. Post, "The AMI meeting corpus," in *Proc. Int. Conf. Methods Techn. Behav. Res.*, 2005, pp. 137–140.
[36] A. Cafaro et al., "The noxi database: Multimodal recordings of mediated novice-expert interactions," in *Proc. 19th ACM Int. Conf. Multimodal Interaction*, 2017, pp. 350–359.
[37] M. Koutsombogera and C. Vogel, "Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus," in *Proc. 11th Int. Conf. Lang. Resour. Eval.*, 2018, pp. 2945–2951.
[38] I. Bhattacharya et al., "The unobtrusive group interaction (UGI) corpus," in *Proc. 10th ACM Multimedia Syst. Conf.*, 2019, pp. 249–254.
[39] C. Wang and G. Chanel, "An open dataset for impression recognition from multimodal bodily responses," in *Proc. 2021 9th Int. Conf. Affect. Comput. Intell. Interaction*, 2021, pp. 1–8.
[40] A. Reece et al., "The candor corpus: Insights from a large multimodal dataset of naturalistic conversation," *Sci. Adv.*, vol. 9, no. 13, 2023, Art. no. eadf3197.
[41] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. v. d. Meij, and H. Hung, "The MatchNMingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 113–130, Jan.–Mar. 2021.
[42] M. Balazia, P. Müller, A. L. Tánczos, A. v. Liechtenstein, and F. Brémond, "Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 70–79, doi: 10.1145/3503161.3548363.
[43] J. C. Coyne, "Depression and the response of others," *J. Abnorm. Psychol.*, vol. 85, no. 2, 1976, Art. no. 186.
[44] C. M. Macovei, Ş. Bumbuc, and F. Martinescu-Bădălan, "The role of personality traits in mediating the relation between fear of negative evaluation and social interaction anxiety," *Front. Psychol.*, vol. 14, 2023, Art. no. 1268052.
[45] M. Asher and I. M. Aderka, "Dating with social anxiety: An empirical examination of momentary anxiety and desire for future interaction," *Clin. Psychol. Sci.*, vol. 8, no. 1, pp. 99–110, 2020.
[46] E. Ermer and K. A. Kiehl, "Psychopaths are impaired in social exchange and precautionary reasoning," *Psychol. Sci.*, vol. 21, no. 10, pp. 1399–1405, 2010.
[47] T. M. Gerlach, R. C. Arslan, T. Schultze, S. K. Reinhard, and L. Penke, "Predictive validity and adjustment of ideal partner preferences across the transition into romantic relationships," *J. Pers. Social Psychol.*, vol. 116, no. 2, pp. 313–330, 2019.
[48] D. Balliet and P. A. V. Lange, "Trust, conflict, and cooperation: A meta-analysis," *Psychol. Bull.*, vol. 139, no. 5, 2013, Art. no. 1090.
[49] J. Berg, J. Dickhaut, and K. McCabe, "Trust, reciprocity, and social history," *Games Econ. Behav.*, vol. 10, no. 1, pp. 122–142, 1995.
[50] N. J. Raihani and P. Barclay, "Exploring the trade-off between quality and fairness in human partner choice," *Roy. Soc. Open Sci.*, vol. 3, no. 11, 2016, Art. no. 160510.
[51] J. F. Rauthmann et al., "The situational eight diamonds: A taxonomy of major dimensions of situation characteristics," *J. Pers. Social Psychol.*, vol. 107, no. 4, pp. 677–718, 2014.
[52] J. Nadler, "Rapport: Rapport in negotiation and conflict resolution," *Marquette Law Rev.*, vol. 87, no. 4, 2004, Art. no. 25.
[53] M. T. Braun, S. W. Kozlowski, T. A. Brown, and R. P. DeShon, "Exploring the dynamic team cohesion–performance and coordination–performance relationships of newly formed teams," *Small Group Res.*, vol. 51, no. 5, pp. 551–580, 2020.
[54] N. Koudenburg, T. Postmes, and E. H. Gordijn, "Conversational flow and entitativity: The role of status," *Brit. J. Social Psychol.*, vol. 53, no. 2, pp. 350–366, 2014.
[55] M. C. Ashton and K. Lee, "The Hexaco–60: A short measure of the major dimensions of personality," *J. Pers. Assessment*, vol. 91, no. 4, pp. 340–345, 2009.
[56] B. E. Hilbig, I. Zettler, F. Leist, and T. Heydasch, "It takes two: Honesty–humility and agreeableness differentially predict active versus reactive cooperation," *Pers. Individual Differences*, vol. 54, no. 5, pp. 598–603, 2013.
[57] L. Peters, M. Sunderland, G. Andrews, R. M. Rapee, and R. P. Mattick, "Development of a short form social interaction anxiety (SIAS) and social phobia scale (SPS) using nonparametric item response theory: The SIAS-6 and the SPS-6," *Psychol. Assessment*, vol. 24, no. 1, 2012, Art. no. 66.
[58] Z. Wei, Z. Zhao, and Y. Zheng, "Moderating effects of social value orientation on the effect of social influence in prosocial decisions," *Front. Psychol.*, vol. 7, 2016, Art. no. 952.
[59] N. P. Li, J. M. Bailey, D. T. Kenrick, and J. A. Linsenmeier, "The necessities and luxuries of mate preferences: Testing the tradeoffs," *J. Pers. Social Psychol.*, vol. 82, no. 6, pp. 947–955, 2002.
[60] A. Pahor, T. Stavropoulos, S. M. Jaeggi, and A. R. Seitz, "Validation of a matrix reasoning task for mobile devices," *Behav. Res. Methods*, vol. 51, no. 5, pp. 2256–2267, 2019.
[61] J. Broekens and W.-P. Brinkman, "Affectbutton: A method for reliable and valid affective self-report," *Int. J. Hum.- Comput. Stud.*, vol. 71, no. 6, pp. 641–667, 2013.
[62] A. Kirmani, R. W. Hamilton, D. V. Thompson, and S. Lantzy, "Doing well versus doing good: The differential effect of underdog positioning on moral and competent service providers," *J. Marketing*, vol. 81, no. 1, pp. 103–117, 2017.
[63] C. W. Leach, N. Ellemers, and M. Barreto, "Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups," *J. Pers. Social Psychol.*, vol. 93, no. 2, pp. 234–249, 2007.

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.

[64] R. E. D. Vries, "The 24-item brief hexaco inventory (BHI)," *J. Res. Pers.*, vol. 47, no. 6, pp. 871–880, 2013.
[65] C. Raman, N. R. Prabhu, and H. Hung, "Perceived conversation quality in spontaneous interactions," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 2901–2912, Oct.–Dec. 2023.
[66] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.
[67] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. 2016 IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.
[68] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSmile: The munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.
[69] B. Schuller et al., "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proc. 17TH Annu. Conf. Int. Speech Commun. Assoc.*, 2016, vol. 8, pp. 2001–2005.
[70] M. Pushkarna, A. Zaldivar, and O. Kjartansson, "Data cards: Purposeful and transparent dataset documentation for responsible AI," in *Proc. 2022 ACM Conf. Fairness, Accountability, Transparency*, 2022, pp. 1776–1826.
[71] M. Kardas, A. Kumar, and N. Epley, "Overly shallow?: Miscalibrated expectations create a barrier to deeper conversation," *J. Pers. Social Psychol.*, vol. 122, no. 3, pp. 367–398, 2022.
[72] T. R. Cohen, T. Wildschut, and C. A. Insko, "How communication increases interpersonal cooperation in mixed-motive situations," *J. Exp. Social Psychol.*, vol. 46, no. 1, pp. 39–50, 2010.
[73] T. M. Hrkalovic, B. Dudzik, H. Hung, and D. Balliet, "Partner perceptions during brief online interactions shape partner selection and cooperation," *PLoS One*, vol. 20, no. 4, 2025, Art. no. e0318137.
[74] T. Fiske, A. J. Cuddy, P. Glick, and J. Xu, "A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition," *J. Pers. Social Psychol.*, vol. 82, no. 6, pp. 878–902.
[75] A. Koutsoumpis et al., "Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning," *Comput. Hum. Behav.*, vol. 154, 2024, Art. no. 108128.
[76] A. Dempster, D. F. Schmidt, and G. I. Webb, "MiniRocket: A very fast (almost) deterministic transform for time series classification," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2021, pp. 248–257.
[77] B. Dudzik, S. Columbus, T. M. Hrkalovic, D. Balliet, and H. Hung, "Recognizing perceived interdependence in face-to-face negotiations through multimodal analysis of nonverbal behavior," in *Proc. 2021 Int. Conf. Multimodal Interaction*, 2021, pp. 121–130.
[78] G. W. Watson, T. Douglas, R. Berkley, R. Madapulli, and Y. Zeng, "Are past normative behaviors predictive of future behavioral intentions?," *Ethics Behav.*, vol. 19, no. 5, pp. 414–431, 2009.
[79] M. F. Jung, N. Martelaro, and P. J. Hinds, "Using robots to moderate team conflict: The case of repairing violations," in *Proc. tenth Annu. ACM/IEEE Int. Conf. Hum.-Robot Interaction*, 2015, pp. 229–236.
[80] B. Dudzik and J. Broekens, "A valid self-report is never late, nor is it early: On considering the" right" temporal distance for assessing emotional experience," 2023, *arXiv:2302.02821*.
[81] N. Koenig-Lewis and A. Palmer, "The effects of anticipatory emotions on service satisfaction and behavioral intention," *J. Serv. Marketing*, vol. 28, no. 6, pp. 437–451, 2014.
[82] M. Tamir and Y. E. Bigman, "Expectations influence how emotions shape behavior," *Emotion*, vol. 18, no. 1, 2018, Art. no. 15.
[83] S. Chen, H. Cheng, and Y. Huang, "Emotion recognition in self-regulated learning: Advancing metacognition through AI-assisted reflections," in *Trust and Inclusion in AI-Mediated Education*. Cham, Switzerland: Springer, 2024, pp. 185–212, doi: 10.1007/978-3-031-64487-0_9.
[84] R. Dazeley, P. Vamplew, C. Foale, C. Young, S. Aryal, and F. Cruz, "Levels of explainable artificial intelligence for human-aligned conversational explanations," *Artif. Intell.*, vol. 299, 2021, Art. no. 103525.

Tiffany Matej Hrkalovic received the BSc and MA degrees in psychology from the University of Zagreb, Zagreb, Croatia, and the second MSc degree in brain and cognitive science. She is currently a postdoctoral researcher with the Jheronimus Academy of Data Science. Her research interests include understanding partner selection for cooperation to create intelligent systems that can support human-human interactions in joint undertakings and human-AI interactions.

Bernd Dudzik (Member, IEEE) is currently an assistant professor with the Department of Intelligent Systems, TU Delft, Delft, Netherlands. His work explores context-sensitive approaches for multimodal modeling of human cognitive-affective processes (e.g., memory recollection or cognitive appraisals) during everyday interactions. His research interests include affective computing and AI-human collaboration. He is also an active member of the AAAC and associate editor for *IEEE Transactions on Affective Computing*.

Daniel Balliet is the founder of the Amsterdam Cooperation Laboratory, and is currently the co-director with Cooperation Databank. His research focuses on understanding human cooperation. He applies experiments, field studies, and meta-analysis to test evolutionary and psychological theories of cooperation. His work addresses issues related to how people think about their interdependence in social interactions, how people condition their cooperation to acquire direct and indirect benefits, and understanding cross-societal variation in cooperation. He was the recipient of an ERC Starting Grant (2015-2020) and ERC Consolidator Grant (2020–2025).

Hayley Hung (Member, IEEE) is currently an associate professor with the Delft University of Technology, Delft, The Netherlands. She also leads the Human Oriented Machine Intelligence Laboratory. Her research interests include in social signal processing, multi-sensor processing, machine learning, and ubiquitous computing, and focuses on devising novel pattern recognition and machine learning methods to automatically interpret group social and affective behavior during face-to-face human interactions.

Authorized licensed use limited to: TU Delft Library. Downloaded on December 08,2025 at 12:19:47 UTC from IEEE Xplore. Restrictions apply.