ISSN 2186-7437

NII Shonan Meeting Report

No. 202

Conversational Qualities in Dyadic and Group Interactions

Shogo Okada
Yukiko Nakano
Wolfgang Minker
Elisabeth André

October 23–26, 2023

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Conversational Qualities in Dyadic and Group Interactions

Organizers:
Shogo Okada (Japan Advanced Institute of Science and Technology, Japan)
Yukiko Nakano (Seikei University, Japan)
Wolfgang Minker (Ulm University, Germany)
Elisabeth André (Augsburg University, Germany)

October 23–26, 2023

1 Description of the Meeting

Automatic evaluation of conversation qualities has become one of the vital points for interactive computer systems (ICCs) of the next generation. Although current ICCs are able to understand what humans are talking about, they are incapable of catching how humans are talking. Moreover, recent research on human-human interaction has shown that more than 90 % of the information contained in speech and visual input is non-verbal. Important parts of this information include conversational qualities (characteristics). While implicitly transmitted among humans during the conversation, they significantly influence the entire conversation and therefore its quality. Modern robotic and computer systems may not even suspect that something is going wrong during the interaction until it is "accidentally" aborted. However, if such systems were capable of catching and analyzing all available conversational qualities shown by human agents, they would act appropriately, embodying naturalness, confidence, and intelligibility. When people talk to each other, they change their verbal and nonverbal communication behaviors according to those of their partner. Therefore, user adaptivity represents an essential issue in improving the quality of human-robot and, more generally, human-agent interaction. There are many potential applications where user adaptivity adds value for enhancing user experience. For example, E-Health is quite relevant as a use case for adaptive human-computer interactions with AI systems since tailored and explainable interventions are needed for long-term engagement. A virtual coach needs to know about the user's personality and cultural background since they impact the user's consideration of health and motivation. The aim of this Shonan meeting is therefore to discuss challenges that may arise in Conversational Qualities Assessment (CQA) during human-human and human-robot/agent interaction. Possible usage of these CQA Systems in various industrial spheres with a focus on application within Human-Computer (Robot) interactions will also be discussed. Finally, a roadmap of the technology development will be established. To move closer to the solutions to the challenges described above, significant

efforts are required from the scientific community involving specialists in the fields of computer science, medicine, and psychology. Sharing and combining expertise from various scientific fields may lead to synergies, which will allow us to create new ways to build solutions or increase the effectiveness of existing approaches. The Shonan Meeting will help to explore possible challenges and jointly develop the respective research fields, whose contributions will serve as a research agenda for main directions. Therefore, we will invite keynote speakers from the respective research fields, whose contributions will serve as a basis for breakout sessions. In these sessions, participants will work actively on specific research objectives in small groups. This will help foster an interdisciplinary understanding and cooperativity. The results of the breakout sessions will then be discussed with the whole plenum. We have decided to structure the workshop into four different areas. These include typical research challenges that need to be addressed when recognizing and evaluating conversational qualities in both dyadic and group interactions. Secondly, probable encountered problems of development, testing, and evaluation will be discussed. Further, the use cases and industrial applications will be covered as well. Moreover, the ethics and societal impact of developed technologies need to be considered as well. To accomplish that, the participants will gather together in three Working Groups (WG) and discuss the aforementioned topics. Each WG will have three working sessions (on October 23rd and 24th) where the WGs can decide for themselves how to divide the topics among the sessions (one session - one topic, each session - three topics, etc.). Every WG will be asked to document their preliminary results using a given template and present them to the other WGs at the end of the first two meeting days. On October 25th, the WGs will be regrouped in a way that each WG is responsible for a single topic, discussing and summing up the outcome of the last two meeting days (WG 1 - Topic 1, WG 2 - Topic 2, WG 3 - Topic 3). Afterward, the final results for each topic will be presented by the respective WG. The WGs are the following:

Research challenges:

- Multimodal CQA in Human-Human Interaction and HRI
  - Influence of paralinguistics on CQA.
  - Multimodal Machine Learning for modeling dyadic and group interaction.
  - Multimodal CQA: approaches and fusion techniques.
  - Online versus face-to-face communication.
- Human-Robot and Human-Human Interaction
  - Human-Computer and Human-Robot Interaction making use of CQA.
  - CQA of Human-robot interaction strategies in terms of adaptivity.
  - Group level performance modeling (Cohesion, Group output).
  - CQA of diverse groups (human-only group, human-robot group).
- Influence of User Properties, Personality and Emotions
  - Personality trait modeling (Personality, Attitude, Engagement, Social skills).

  - Modelling of the influence of individual differences (age, gender, and language) on multimodal interaction.
  - Affective modeling on multimodal interaction (Sentiment, Empathy).
  - The role of emotions in human-robot conversations.
- Resources and Data for CQA
  - Data collection: setup and choice of sensors.
  - Available Tools for data collection, processing, and annotation
  - Data robustness: suppressing noise, learning from small data, and handling missing data.
  - Data quality: Reliable annotation and motivation of participants.

Development, testing and evaluation & Ethics and societal impact:

- Experimental design, user studies, and evaluation of systems for automatic CQA.
- Experimental set-up accounting for real-world, real-time, large-scale conditions. Engineering approaches to CQA: life-cycle, requirement elicitation, robustness to change, standardization, and simulation.
- Development models, tools and strategies: middleware, languages.
- Meaningful and explainable system evaluation.
- Description, development, and sharing of resources: corpora compilation, annotation tools, and approaches, crowdsourcing approaches.
- Sensing devices and frameworks designed for CQA.
- Competitive research challenges planning and organization.
- Data protection and privacy by design and default.
- Legal issues.
- Trust and usability.
- Social responsibility.
- Interdisciplinary approach by integrating findings of physiology and social science

Use cases, prototypes, and industrial applications:

- Computer-mediated human-to-human interaction.
- Health applications
- Social Human-Robot Interaction.
- Social skill training applications
- Tutoring and E-learning applications.

- Coaching applications.
- Energy, memory, and computing efficient CQA. Model pruning/shrinking for usage on portative devices.
- Success stories, functional systems, and industrial challenges.

2 Meeting Schedule

Conversational Qualities in Dyadic and Group Interactions

Time Table	Arrival Day	1st Day	2nd Day	3rd Day	Final Day
	October 22nd (Sunday)	October 23rd	October 24th	October 25th	October 26th
7:00 - 7:30					
7:30 - 8:00		Breakfast	Breakfast	Breakfast	Breakfast
8:00 - 8:30					
8:30 - 9:00					
9:00 - 9:30		Introduction	Keynotes: Catherine Pelachaud (Topic 1)	Working Groups - Session 5 (Topic	Working Groups - Elaboration of final
9:30 - 10:00		Pecha Kucha 1	Patrick Gebhard (Topic 3)	-related)	results
10:00 - 10:30		Pecha Kucha 2			
10:30 - 11:00		Pecha Kucha 3	Break + Photo Shoot	Break	Presentation of final results
11:00 - 11:30		Break	Kristiina Jokinen (Topic 2)	Working Groups - Session 6 (Topic	Wrap up and Farewell
11:30 - 12:00		Keynotes: Albert Ali Salah (Topic 1)	WG - Session 3	-related)	Lunch
12:00 - 12:30	Early check-in (negotiable)				
12:30 - 13:00		Jean-Claude Martin (Topic 2)	Lunch	Lunch	
13:00 - 13:30		Lunch	Keynotes: David Traum (Topic 1)		
13:30 - 14:00			Martin Baumann (Topic 2)	Excursion:	
14:00 - 14:30		Shiro Kumano (Topic 3)	Gavin Doherty (Topic 3)	Visiting Jomyoji and Hokokuji temples	
14:30 - 15:00		WG Allocation	Break	(with Japanese Tea ceremony)	
15:00 - 15:30	Regular check-in	WG - Session 1	WG - Session 4		
15:30 - 16:00					
16:00 - 16:30					
16:30 - 17:00		Break			
17:00 - 17:30		WG - Session 2			
17:30 - 18:00			Discussion of Results		
18:00 -18:30		Dinner	Dinner		
18:30 - 19:00					
19:00 - 19:30	Welcome Banquet			Banquet	
19:30 - 20:00					
20:00 - 20:30					
20:30 - 21:00					
21:00 - 21:30					
21:00 -22:00					

3 Working Groups

The participants of this meeting were divided into three Working Groups (WG) and discussed the following three topics.

- Research challenges
- Development, testing, evaluation & ethics and societal impact
- Use cases, prototypes, and industrial applications

Overview of Talks

Multimodal Analysis of Dyadic Interactions in Mental Healthcare

Prof. Dr. Albert Ali Salah, Utrecht University, Netherlands

Affective and social computing has enabled novel ways of analysis of human behavior at different scales. A large range of potential applications are currently being investigated, and some are already commercialized. From automatic transcription and logging of verbal and non-verbal behavior to early diagnosis and monitoring of patients, multimodal technologies show great potential. Mental healthcare is one of the major application areas for conversational agents, as a vast majority of the world's population does not have access to a qualified therapist. To implement successful agents, both analysis and synthesis of realistic and relevant behaviors are important. In my talk, we will focus on the analysis. In particular, a sub-class of analysis approaches concern dyadic interactions, where the interpretation of social signals becomes very important. we will discuss research challenges and opportunities in this area by giving examples from two specific problems we have been working on recently; namely, the analysis of therapist-child interactions in child play therapy interventions, and the analysis of therapist-patient interactions during psychotherapy sessions. In both cases, we work with loosely controlled recording conditions and test a range of multimodal approaches for extracting social cues, such as the expressed affect of the participants, and the working alliance between the therapist and the patient. We will point out the capabilities of the tools we currently have for each modality, and in what ways they need to be developed further.

Social and Motivational Interactions

Prof. Dr. Jean-Claide Martin, CNRS-LIMSI, Paris-Sud University, France

The research of my lab has two goals: (1) gaining a better understanding of human interactions and (2) designing motivational and social human-machine interactions applied to health and disability. To this end, we study and adapt psychological theories of motivation, personality, social interaction, and multimodal communication. These theories are used to design and evaluate personalized and motivational human-machine interactions integrating animated virtual agents and personalized interactive mobile applications. In the field of health and disability, we work with hospital specialists in various pathologies such as low back pain, diabetes, and Alzheimer's disease, as well as with care facilities (intellectual disability, autism). Those problems require a multidisciplinary approach and are based on cooperation with researchers in human-computer interaction and psychology, doctors, disability specialists, and industrial partners. In this speech, we will focus on two complementary areas: personalized social skills training (for disabled people, professionals, or the general public) and personalized motivation for behavior change and health as well as designing models (normal and pathological models of emotion evaluation), software

tools (MARC virtual agent platform, automatic analysis of social signals), and evaluation methods (e.g. evaluation of behavioral models).

More Statistical Treatments of Qualitative Self-reports in Conversations: Response Style Removal and Model Training and Evaluation Under Aleatoric Uncertainty

Prof. Dr. Shiro Kumano, Nippon Telegraph and Telephone Corporation (NTT), Japan

Measuring the quality of conversations can be approached in at least three ways, similar to emotion measurement: subjective experience, physiological responses, and behavior. A comprehensive measurement and analysis that integrates these approaches would be vital. While there have been significant advancements in natural language processing for linguistic reporting, behavioral measurement techniques, and the development of physiological response measurement devices, the most direct and straightforward method, specifically quantitative self-report using tools like the Likert scale and visual analog scale, still faces challenges in terms of reliability. These challenges include subjective biases and reproducibility. To address these issues, leveraging the wisdom of the crowd, it's common to employ multiple external annotators and average their assessments. However, this approach is fundamentally suited only for constructs perceived by individuals outside the conversation. It becomes challenging to apply to constructs that the target person genuinely experienced. The challenge remains even for perceived states when evaluators are limited to the interlocutors. In this talk, the approaches to alleviate the bias and reproducibility issues in Likert-scale data will be introduced, including statistical models designed to remove response styles from observed data, and a method for jointly training and evaluating models that handle data with aleatoric uncertainty. Hopefully, this talk will inspire discussions during the meeting about the types of data to be collected and how they should be analyzed.

Building and Evaluation Adaptation Mechanisms for Socially Interactive Agents

Prof. Dr. Catherine Pelachaud, Centre National de Recherche Scientifique-Institut de Systèmes Intelligents et de Robotique (CNRS-ISIR), Sorbonne University, France

In this talk, we will present our effort in building a Socially Interactive Agent (SIA) able to interact verbally and nonverbally with human interlocutors. SIA has been endowed with the capacity to display a wide range of communicative and emotional behaviors. During an interaction, humans continuously adapt their behaviors at different levels involving formality of language, imitation, synchronization, etc. As a partner in an interaction, SIA has to engage with its human interlocutors. It requires managing turn-taking but also adapting to its multimodal behaviors and conversational strategies. To this aim, we have

developed several adaptation mechanisms where the SIA can adapt its conversational strategies or its multimodal behaviors. These models drive the behaviors of these agents. They were evaluated through perceptive studies where human participants interacted with them in real-time. We will present these different adaption mechanisms, the architecture of the human-agent platform in which they are implemented, and the evaluation studies we conducted.

Exploring the Individual Emotional Experience in Dyadic Interactions between Humans and Machines

Dr. Patrick Gebhard, German Research Center for Artificial Intelligence, Germany

Several aspects influence the quality of a conversation in human-machine interaction. The individual emotional experience has a significant impact on the assessment of the quality. Therefore, this talk starts with investigating the functions of emotions, relevant theories of emotions, and how this can be computationally modeled and realized in systems with socially interactive agents. Central relevant psychological research processes are also addressed and illustrated with examples showing approaches and a peek into relevant evaluation methods.

Challenges in Evaluating Conversational Qualities for Trustworthy HRI Applications

Prof. Dr. Kristiina Jokinen, AIRC AIST Tokyo Waterfront, Japan

Although LLMs have been available for some time and quietly made their progress in various scientific fields, the launch of ChatGPT at the end of 2022 took the world by surprise and awe, and also made the public aware of the power of generative models. Also in spoken dialogue research, this supported a rapid paradigm shift from the traditional dialogue structures, standards, and practices to design and development via prompt design and smoothly going chatty conversations.

However, several known issues concern LLMs and their use ranging from false information to ethical concerns. In practical applications that do not only aim at engaging chats but at providing trustworthy information and good services, it is important to focus on evaluation methodologies and re-think such concepts as engagement, user satisfaction, reliability and truthfulness, long-term interaction, privacy, and ethics.

In this talk, I will discuss challenges in human-robot interactions and focus especially on the evaluation of language-capable robots. I will draw examples from my existing projects.

What are (non-)Conversational Qualities in Group Interaction?

Prof. Dr. David Traum, Department of Computer Science, University of Southern California, USA

The keynote speech delves into the core question: "What defines conversational qualities, and how do they differ from non-conversational aspects?" We will examine various interactive activities, breaking down their levels of conversationality. Recent dialogue systems will be analyzed to pinpoint elements contributing to conversational or non-conversational interactions. The speech concludes by speculating on the evaluation of conversational qualities, aiming for a straightforward exploration of this intricate subject.

Ethical and Psychological Challenges in Human-Technology Interaction

Prof. Dr. Martin Baumann, Ulm University, Germany

Technological progress in recent decades has led to the development of human-machine systems in which automation is increasingly able to make decisions independently and also carry out the corresponding actions autonomously. The automation thus became an at least partially autonomous agent in these human-machine systems and by this also in the context that contains these human-machine systems. While such systems are supposed to provide enormous potential to increase safety, efficiency, and comfort, this potential can only be fully exploited if human users sufficiently understand these systems, are able to predict their future behavior and develop an appropriate level of trust in them. However, this can only be achieved if these systems not only act reliably but also possess the ability to act as supportive and cooperative partners to humans. That, is these technical systems must have the ability to communicate in an appropriate and efficient way to make their current state, plans, goals, and their current understanding of the situation transparent and easy to understand for their human cooperation partners. At the same time, these systems must possess some information about their human cooperation partners, their goals, current state, and situation comprehension to be able to create and support the generation and maintenance of a shared understanding of the situation between the humans and the technical systems and on this basis to provide appropriate and timely support.
In this presentation, some results of recent research projects on different aspects of cooperative human-machine interaction are shown and their possible implications for the design of cooperative automation are discussed, especially with reference to conversational qualities.

Towards More Personalized Delivery of Digital Mental Health Interventions

Prof. Dr. Gavin Doherty, School of Computer Science and Statistics, Trinity College Dublin, Ireland

Using digital mental health technology is inherently a sensitive and private experience. Despite sustained interest in the possibility of in-the-moment interventions leveraging the sensing capabilities of mobile phones and wearable devices, and increasing sophistication of conversation-based interaction, currently, available systems do not necessarily deliver a personal and personalized experience. In this talk, I will discuss our efforts to support more personal and tailored experiences for the users of digital mental health interventions, looking at the design space of these interventions, and considering motivations, feasibility, and acceptance. The talk will particularly consider possibilities for the integration of Machine Learning and Affective Computing capabilities within mental health interventions and some of the difficulties surrounding these. These applications may be incorporated into systems used by clinicians, to support clinicians in their interactions with clients, and with the clients themselves.

4 Summary of Discussions

This section shall give an overview of the very fruitful and wide-ranging discussions during the Meeting.

The field of Conversational Quality Analysis (CQA) faces significant research challenges in both dyadic and group interactions. These include representing context features, creating standardized assessment instruments, analyzing online conversations, and addressing data-related issues such as quantity, quality, annotation, and privacy. Modeling challenges involve salient modality selection, incorporating theory of mind, and developing evaluation systems. The development, testing, and evaluation of CQA systems require carefully designed experimental setups that consider factors such as interaction goals, domain specificity, and methodological paradigms. Evaluation methods encompass both subjective and objective measures, with a need for standardized scales and corpora specifically tailored for group interactions.

CQA applications span various domains, including health, education and training, group collaboration, tele-customer services, and computer games. In healthcare, CQA is being used for telemedicine, mental health interventions, and analysis of patient-clinician interactions. Educational applications include job interview training, public speaking coaching, and social skills development. Group collaboration tools leverage CQA to enhance creative communication and decision-making processes. In customer service, CQA is employed to improve chatbot interactions and handle complaints more effectively. The gaming industry utilizes CQA to enhance non-player character interactions and support teamwork in multiplayer environments.

The development and implementation of CQA technologies raise significant ethical and societal concerns. These include potential privacy violations, manipulation of conversation qualities, psychological harm, and the perpetuation of societal biases. However, CQA also offers positive outcomes such as enhanced public awareness of nonverbal communication cues and early detection of biases in critical systems. Addressing these issues requires compliance with ethical and legal standards, responsible design practices, and increased transparency in CQA systems. Future work in this field should focus on integrating CQA functionality into real-world applications while carefully considering the ethical implications and societal impact of these technologies.

In the following subsections, we separately go over the main topics of the successfully accomplished Shonan Meeting. Firstly, we define so-called Conversation Qualities, then discuss the Research challenges in their assessment, afterward, we describe the development, testing, and evaluation of CQA frameworks and provide an overview of the use cases, prototypes, and industrial applications of such systems. Finally, we discuss the ethics and societal impact of leveraging CQA systems in human society.

4.1 Conversation Quality Definition

Different concepts are meant by this term. For example, dictionary.com says "(in public speaking) a manner of utterance that resembles the spontaneity and informality of relaxed personal conversation."[1] This refers to the properties of

[1] https://www.dictionary.com/browse/conversational-quality

speech that make it seem like the speaker is in a conversation. Other sources provide the following definition: "Conversational qualities refer to the attributes and characteristics that contribute to effective and engaging conversations between people. These qualities are essential for clear communication, building rapport, and ensuring that interactions are meaningful and productive." This focuses on the qualities of conversation itself (e.g. good or not so good). Another definition used by the workshop organizers describes Conversational Qualities (CQs) as referring to a broad spectrum of non-linguistic attributes and characteristics that play a significant role in dyadic and group interactions, shaping the dynamics and effectiveness of communication. These qualities encompass various aspects of interpersonal exchange, including affective states, non-verbal communication, turn-taking, social dynamics, cultural sensitivity, empathy, and others. This focuses on the observable differences in these features based on different kinds of settings, such as differences between dyadic conversations or multi-party conversations.

Conversational qualities can be analyzed from different perspectives. First and foremost, from the sender and receiver's point of view, information content contributes to them. Second, mechanics of turn-taking, and repair (i.e., timing, feedback channeling; classical concepts from conversation analysis) can be considered as factors. From the (multi-party) conversation, the interaction dynamics and social multi-level adaptation are further contributing to the CQs.

For different conversational contexts, different features may be considered to be more prominent. In some cases, the identification of what is relevant for a particular context may be a research challenge in itself. In the measurement of CQ, we distinguish between high-level and low-level features. For example, there has been a lot of research on leadership in groups [30, 36, 4], which can be considered a high-level construct, and low-level features like direct gaze [17] and interruptions [19] are known to be relevant to that particular construct.

In conversation analysis, the theory of mind plays an important role [25, 14]. Who knows what in the group, and how does each participant model that (imperfectly), who contributes to the group goals in what way, and when, are important questions. This requires a multi-layered representation for relevant features that quickly grows in complexity as the group size is increased. Subsequently, there is a need for sets of features that are reliably and objectively extracted from a group conversation, which can then be further processed to gain insight into higher-level constructs with more room for subjectivity.

High-level features are perception level, subjective qualitative concepts that include:

- affective states (emotions, mood);
- empathy, active listening;
- cohesion, atmosphere, conviviality;
- synchrony, alignment, entrainment, rapport, mimicry and imitation;
- engagement;
- dominance distribution;
- trust, honesty, and openness;

- social attitude;
- naturalness, smoothness;
- group purpose (which may be different from individual goals). It can be external (building something) or consolidating members (group therapy).

Low-level features, however, are observable, measurable characteristics of the conversation, including:

- paralinguistics and facial expressions, touch, bodily expressions;
- affective primitives (arousal, valence, dominance);
- response speed and timing;
- speaking rate, intonation;
- gaze, proxemics, group formations in space;
- breathing patterns, heart rate, brain waves;
- gestures;
- syntactic and grammatical language use, morphological distribution;
- and overall language use.

These qualities can also be categorized into individual and group-level features. While individual features can also be studied at the group level, the other direction is not possible. For example, affective states can be both an individual and a group-level feature [31], however, engagement, proxemics, and cohesion can be studied only at the group level. The computation of high-level functions may involve different participants in a group. For example, task cohesion, and task performance can be computed with the active members of the group, while social cohesion and conviviality may be computed with all the members of the group.

4.2 Research Challenges

4.2.1 Main Differences Between Dyadic and Group Interactions for Conversation Qualities Assessment

CQA differs largely on a range of aspects between dyadic and group interactions. Below we elaborate on these aspects one by one.

Social and Temporal Dynamics. In a dyad, one is generally either speaking or listening and there is little distinction between ending one's turn and assigning the turn to another participant. In a group, there are different kinds of actions to assign a turn or request a turn [32]. The timing of interactions can vary greatly in group settings due to the presence of multiple participants, making modeling and predicting behaviors more challenging, especially considering that subgroups may be formed dynamically. Group conversations, even if intended to be more structured compared to intimate dyadic interactions, naturally contain more noise, speech overlap, and body occlusion/face pose for processing the low-level features [1, 40].

In a group setting, attention distribution and engagement of individuals varies compared to dyadic interactions. For instance, one might assume strong membership and mutual engagement in a dyadic setting, but in a group, participation might be more spread out. In a dyadic setting, the addressee is clear, whereas in a group setting, the addressee can be an individual, a subset of participants, or all non-speaking participants and it is not straightforward to assess the addressee.

The type of conversation changes the dimensions we should analyze synchrony or group behavior. In a dyadic interaction, one can analyze alignment or mirroring, while in a group setting, we need to be concerned with both modeling some group notions (e.g. emotional contagion, and perceived group consensus), as well as alignment to any of the individuals.

Hierarchy and Role Dynamics. Hierarchies and roles are more complex in groups. In dyadic interactions, the roles are generally more clearly defined [29]. The existence of sub-groups may have an effect on task cohesion (performance), but also on social cohesion/conviviality.

Other aspects where dyadic and group CQA differ include mechanisms for joining and leaving an existing conversation, proxemics (where to stand in relation to others, F-formations), the perception of in-group and out-group, and processing speech and gaze direction. This breakdown highlights the inherent complexities of group interactions versus dyadic ones, especially when considering machine learning models and robotic applications.

Non-human Participants. While a robot or virtual agent (VA) in a dyadic interaction needs to adapt to one human [27, 46], in a group setting, the decision becomes complex – should the system adapt to an individual, a subgroup, or the entire group? We see in the literature that non-human participants may assume different roles in groups [38, 45]. An ordinary participant role is possible for flat group structures; for instance, a non-human player in a gameplay setting. Other common roles are facilitators and teacher/guide roles, which assume a non-symmetrical relationship between participants. An autonomous agent equipped with many sensors and a lot of processing power may be in an ideal position to monitor signals from the rest of the group.

When there is high-quality sensor data (i.e. high spatio-temporal resolution), autonomous systems may be even more accurate than humans in classifying social signals, but research in the ideal lab settings may not easily generalize to more "in the wild" settings. Furthermore, there are (often culture-dependent) a-priori effects and novelty effects in the interactions with autonomous agents.

Infrastructure. In a physical group setting, we may need more sophisticated equipment (e.g., sensor arrays) to collect, manage, analyze, and react to the participants' signals [18, 16, 8]. Multiple cameras will typically be used to be able to capture facial expressions and gaze features in a group setting. Wearable sensors, such as sociometric badges, can provide proximity, and low-level features, that are intuitive, easy to process, and less intrusive in sensitive settings (such as mental healthcare scenarios). The infrastructure is different for virtual groups in comparison to physical groups.

The dynamics of group interactions can change significantly in online settings. For instance, latencies in online communication can affect group dynamics, and the potential of turning off some sensors changes interaction patterns. The equipment of the group members (e.g. cameras and microphones) may be of different quality and the low-level feature distribution may be affected by this

(e.g. some subtle emotions may be impossible to detect on low-level sensors). This may introduce biases into the analysis, which may require pre- or post-processing for more reliable results. Missing data also has a different nature in online settings, where it is perfectly possible for a participant to mute its sensors for periods at a time, during which it becomes impossible to estimate engagement and attention. On the other hand, online settings enable additional modalities (such as a meeting chat), where asynchronous information can pass between participants.

From a computational infrastructure perspective, group analysis requires more resources than dyadic interaction analysis. The potential interaction of modalities in dyadic conversation scales quadratically with the number of modalities. However, for larger groups, the potential dyadic sub-interactions scales exponentially with the group size. In group interactions, more spatial organization settings and more data channels exist, and there may be irrelevant interactions that need to be discounted. All these cause the group setting to be much more difficult than the dyadic setting.

In groups, there is a greater variation of signals, simply due to the presence of more sources. In socially and/or culturally mixed groups, non-verbal cues may be more difficult to interpret, and there is more room for mutual misunderstanding of cues. Address resolution and diarization are more complex than in dyadic interactions. Furthermore, normalization of signals may become an issue.

4.2.2 Biggest Research Challenges in Conversation Qualities Evaluation

As CQs are very complex and comprehensive, the modern scientific community faces many challenges trying to build automatic CQA systems. Below we provide the biggest challenges and possible ways to solve them:

How to Represent Context Features in which CQ is Evaluated? CQs gain meaning within a given context. Context refers to characteristics that are relevant for interpreting, deliberating on, and behaving within an interaction, such as prior situations and anticipated future situations. Some aspects do not change or slowly change during the interaction. Others change more frequently. Some can be directly influenced by the interaction (such as the topic or social obligations of participants) while others can only be observed (such as the locations of other individuals and objects). Aspects of context include but are not limited to age, gender, personality, sociocultural background, roles and goals of the participants; social situation and its challenges/relationship, main goal of conversation (negotiation, command, chatting), environment/sphere of communication, social obligation/constraint, physical settings, previous conversations, the sentence a word is used in or the topic(s) of past and present discussion.

Creation of an Instrument for CQA. There are no widely used instruments to measure CQA. Creation of a broad instrument, to be complemented with domain-specific items as required, will be helpful in assessment, and can also lead to the guidance of automatic assessment approaches and adversarial learning.

To produce such an instrument, it may be possible to follow a structured approach, adapted to the specific question. An example was presented by G. Moore and I. Benbasat in [28], which developed an instrument for measuring the

perceptions of adopting an information technology innovation. Their approach started by making a list of relevant items for measuring CQA. The item creation included culling of items that were useful for a very narrow set of use cases. For instance, a very specific conversational context, like a job interview, may have specific quality items that do not have an important function in other settings. Another example is the Godspeed questionnaire by Bartneck [5].

The way the item list was culled was to consult a small set of experts, and each respondent was asked their level of agreement (7-point Likert scale) to include an item or not in the final list. Once the item pools were generated, redundant or ambiguous elements were eliminated.

The second stage is Scale Development, to assess construct validity and to identify any remaining ambiguities. In an initial stage, a set of judges were asked to provide labels for the constructs, without being told what labels were used in the first stage of Item Creation. Consistent placements of items into similar constructs helped with validity. Then, a range of representative judges sorted the items, and the inter-rater reliabilities were assessed via statistical approaches. In several successive sorting rounds, the constructs were refined into a consensus, and processed for conciseness. The third stage is Instrument Testing, which for CQA, would involve a wide range of applications to test the instrument, first with pilot tests, and then with full field tests.

Conversational Quality Analysis in On-line Conversations. Some aspects of conversation are similar across different modalities, such as face-to-face in person vs telephone or video conference, however, some differences in media affordances create differences in conversational qualities. Sometimes these differences can be beneficial, e.g., in making it easier to monitor who is saying what. But sometimes it can make it harder to have high-quality conversations or can otherwise change the nature of the interaction to be less conversation. For example, a short time lag can make smooth turn-taking almost impossible. It is also more difficult to recognize the target of the gaze of other participants in online interactions.

Data and Data Processing Related Challenges. Here, several sub-challenges can be highlighted:

- **Data Quantity, Quality, and Annotation**. One of the leading problems is the lack of rich, sufficiently annotated data. This involves the way the data should be annotated such that we have "Ground-Truth" data. Is it going to be real-time annotation, post annotation by the participants, or by other people (e.g., professionals)? We should further analyze the aspects for which the ground-truth data are helpful (e.g. it could be used to train CQA modules, but perhaps not for creating a conversational agent). Lastly, inter- and intra-person variation in behavior - "conversational" behaviors for one (at one time) might not be "conversational" for another. To alleviate these problems and privacy-related issues, the generation of realistic multimodal data is a need and remains a challenge to be addressed.

- **Synchronization.** A related challenge is the quality and synchronization of the data. There is a time lag occurring in online conversations. It is like each participant is interacting with the other people in a parallel world. How to recover the temporal relationships of the behaviors between the participants for analyzing CQA can be a challenge. Moreover, as a data

preprocessing stage, significant energy is devoted to synchronizing the modalities for further modeling.

- **Privacy and Fairness.** Ensuring the privacy of the data and that it is free of annotation bias is an important challenge. These should be handled at preprocessing, feature (signal) representation as well as modeling stages.
- **Measuring context.** If it is not explicitly given, identifying and measuring the dimensions of context remains a challenge. Measuring the alignment/contagion in a group; dialog and knowledge distribution (who knows what, who contributes to what, and when), and correlating the observable measures with subjective information are other challenges from measurement.
- **Handling Interference.** At all levels of data processing and modalities, we face interference issues. These include but are not limited to occlusions, speech overlap, and background noise that need to be handled.
- **Expertise in Data Processing.** CQA emphasizes non-verbal factors over linguistic data, requiring expertise in newer technologies that are less mature than the older text-based techniques.

CQ Modeling Related Challenges. For this challenge, there are also several challenges to be addressed:

- **Salient Modality/Signal Selection for Fusion.** In most cases different signals are more informative about some qualities vs others. In this case, obtaining a full picture of conversational qualities with only a subset of modalities and deciding on the optimal way/level these modalities will be fused remains as an important modeling challenge.
- **Theory of Mind.** Theory of mind probably has a relationship with the CQA module, and model of others. Connection to the psychology area is very important.
- **Explicit vs Implicit.** We can model the conversation with explicit/implicit models for conversation quality assessments. If the system has a CQA module, the model behavior will be easy to capture. Recent end-to-end E2E modeling by neural networks is another solution, we still can insert a CQA model to the E2E architecture, for example, adversarial learning or reinforcement learning, it will contribute to the model behavior, and also investigate what we are doing in CQA.
- **Context.** We can use multiple approaches to address the problem of context. One of these approaches is controlling the dimensions of context which are relevant to study the conversation qualities you are interested in. Controlling and reducing the dimensions allow for a comprehensible environment. Automatic assessment of context dimensions via machine learning models can assist in further processing. The relative and normative view of interactions also influence the context, so account differences have to be taken into account as well. Those differences include factors like role, age, culture, education, and politeness.

- **Evaluation System.** Evaluation systems are required. There are two main systems used offline: on-the-fly computation and offline evaluation. On-the-fly communication between human-agent interaction requires real-time systems, and there are many challenges; it is hard to make up the evaluation numbers. Off-line is based on corpus analysis, for example. Off-line will be easier but requires careful design of the evaluation system. New evaluation systems can be considered, such as using brain signals or subjective reports as the ground truth.

- **Measurements.** There is a need for different measurements regarding dyadic and group interactions:

  - Coherence/alignments: In the dyadic interactions, coherence or alignments are equal to the individual models. However, in the group interaction case, it is required to consider group sentiments, conflicts, coherence, or alignment.
  - They can be measured with both verbal and paralinguistic features.
  - Interpretability/explainability: extracting observable affective/behavioral cues from multimodal signals and verbalizing them is important, especially in the age of LLMs.
  - Partial privacy: partial (data) removal or transformation of signal (e.g., face, speech) with minimal information loss.
  - Partial privacy/fairness: obtaining sensitive-attribute (age, gender, ethnicity) independent uni- and multi-modal feature representations.
  - Handling gaze: the target of gaze behaviors is difficult to figure out in an online and multiparty conversation.

- **Features.** The CQ modeling challenge can be solved in several ways from the features utilization side. Constant monitoring of dynamic changes using low-level features forms the foundation for higher-level analysis, where individual-level outcomes from machine learning models can be combined to assess group-level features. Alternatively, deep learning models trained on large datasets may extract meaningful patterns directly from raw data. To manage complexity, a divide-and-conquer strategy was proposed, focusing on identifying and addressing issues within subgroups before scaling up. The seminar also emphasized looking beyond traditional machine learning methods, suggesting the use of visualization techniques to leverage human pattern recognition abilities. Addressing the critical issue of data sharing, federated learning was presented as a potential solution to enable broader data utilization without compromising privacy. Finally, the importance of extensive field testing was underscored, emphasizing that some problems require large-scale implementation to properly evaluate performance and develop sustainable financial models, necessitating thoughtful design work from the outset.

4.3 Development, Testing, and Evaluation

4.3.1 Experimental Setups Design for Conversation Qualities Assessment in Dyadic and Group Interactions

Designing experimental setups for CQA in dyadic and group interactions requires a nuanced understanding of several factors. Initially, one should determine the goal of the study and the type of conversations they are analyzing. The aim could be exploratory, seeking out which features are most relevant, or it could focus on specific experiments testing hypotheses about the impact of conversation qualities. Furthermore, it's essential to identify the topic or goal of the conversation, as different interaction activities possess distinct norms. For instance, a brainstorming session may prioritize immediate positive emotions [9], while a debate might be more accepting of negative responses [20, 7] depending on the long-term outcomes of the stated goals.

The domain within which the conversation occurs also plays a pivotal role, and one must consider multi-floor conversations. Deciding on the method involves determining the desired level of control over the environment. Some might prioritize a controlled environment to test specific hypotheses, while others might prefer observational studies that capture natural conversational behavior. This decision also touches on whether to opt for a cross-sectional or a longitudinal study and how much of the interaction can be pre-defined without disrupting its conversational nature. This could involve decisions on strategies like using a confederate or a "Wizard of Oz" approach, establishing turn-taking rules, defining user roles, and more.

There are various methodological paradigms to choose from, including laboratory studies and observational studies. Within laboratory studies, researchers need to decide between within-subject or between-group configurations, as well as whether to use independent samples or round-table pairing and grouping methods. On the other hand, observational studies can be conducted in natural settings or via online platforms like video conferences.

Another dimension to consider is the use of simulation environments. These environments, equipped with socially interactive agents, can be employed to test hypotheses on the predicted course of a conversation after an intervention or to understand the emergent properties when the group size varies. Within this, choices between the Wizard-of-Oz paradigm and functional prototypes will arise.

Furthermore, the experimental design must take into account different group configurations. This includes deciding whether to use similar setups for both dyadic and group interactions and understanding the potential challenges of such an approach. An essential aspect here is the decision on data collection, which could focus on individual participants or the group as a whole.

Finally, while selecting and setting up sensors to capture interaction data, a balance between unobtrusive and obtrusive methods is vital. This choice, while essential for gathering accurate data, should also respect ethical considerations, ensuring the privacy of participants and avoiding potential harm. The scalability and reproducibility issues due to the complexity of the sensors used also demand careful attention.

4.3.2 Evaluation Methods

When evaluating participants' CQs, it's essential to distinguish between subjective and objective measures. Subjective measures include self-report questionnaires, annotations by experts, and feedback from partners. The timing of these measures can vary, ranging from real-time annotations during the conversation to post-annotation assessments. There's also a need to consider the resolution of the evaluations, whether they are frame-based, focused on individual utterances or acts, or based on entire conversational turns.

Objective measures, on the other hand, assess direct responses to conversations, such as turn-taking behaviors, eye gaze, speech amounts, and facial action units. Physiological responses can also be used as objective measures, including brain signal monitoring and autonomic nervous system-related signals like heart rate and Electrodermal Activity (EDA) or Galvanic Skin Response (GSR).

Both quantitative and qualitative methods have their places in CQA evaluation. Qualitative methods include techniques like guided introspection interviews where participants watch videos of their interactions, expert interviews, and "thinking out loud" methods. There's also interest in using Large Language Models (LLMs) as evaluation or analysis tools, though concerns about their controllability need to be addressed.

Quantitative methods emphasize post-hoc and a-priori evaluations. One challenge here is the lack of established evaluation methods specifically tailored for group interactions. Furthermore, there's the added complexity of using LLMs not just for development but also for testing and evaluation.

When it comes to scales for evaluation, various standardized questionnaires can be employed. These include those focused on mental health, like the BDI [6] or PHQ8/PHQ9 [24, 23] for depression, and others related to conversational dynamics, such as the SPRS or SCC. Some scales are designed to gauge the impact of conversations by measuring changes before and after interactions. For studies conducted in natural settings, the Ecological Momentary Assessment [39] method can be beneficial, allowing for real-time evaluations.

Moreover, the use of dyadic and multi-party dialogue corpora can offer a rich source of data for evaluation. However, as LLMs become more widespread in use, new challenges emerge for evaluation methodologies. LLMs are known for their tendency to produce false information [33], so current research is actively exploring methods to fine-tune these models and address potential issues. This exploration includes looking into novel methodologies to employ LLMs effectively in testing and evaluating experimental results.

4.4 Use Cases, Prototypes and Industrial Applications

In the application context, technology-supported or computer-mediated human-to-human interaction enhances communication. Health applications are innovating patient care, while social Human-Robot Interaction is redefining our relationship with machines. Social skill platforms and E-learning tools are reshaping education and training. The landscape is rich with systems supporting dyadic or group interactions but most of these systems have only limited abilities to perform automated CQA.

CQs comprise low-level as well as high-level qualities. Low-level CQ refers to the observable, measurable features. The list of them as well as the definitions

are presented in Subsection 4.1. From observed low-level CQs one may derive 2nd order or high-level CQs (also presented in Subsection 4.1). From inferred high-level CQ a system may derive further information, such as:

- personality traits and social attitudes of participants
- social relationship among participants
- group purpose and task
- trust, rapport
- efficiency and effectiveness of a group conversation with regard to purpose and task
- and many others.

As conversations vary with regard to the roles and goals of the participants, the number of participants, and the situation and context in which they take place, an automated approach to CQA may give different weights to the above-mentioned high-level and low-level qualities. While most conversational systems comprise recognizers of individual qualities, related to CQ, there aren't general CQ-empowered systems yet.

4.4.1 Application Domains

The application of CQA within particular domains can be broadly classified in terms of explicit training, shaping the quality of interactions between conversation participants (such as providing real-time feedback), and specific tasks such as assessment. In the following paragraphs, we describe various application areas of CQA in such fields of human activity as Health, Education and Training, Group Collaboration, Tele-customer Services, and Computer Games.

Health. Given the growth in telemedicine and digital health applications in recent years, there are many opportunities for the use of CQA both in face-to-face and computer-mediated healthcare settings. Beyond dyadic/group and in-person/remote, a key distinction is the person to whom the output of the CQA is being provided, whether this is provided only to the clinician, for the purposes of training or informing their professional practice, or whether it is also available to the client/patient in order to provide them with insight. There may also be issues of transparency and power imbalance to be considered where it is only provided to the clinician. Arguably the introduction of this technology would shift control over the interaction further in the direction of the clinician.

There are a number of possible uses for the analysis of conversations between patients and healthcare professionals, including psychotherapy and counseling. Examples would include analysis of empathy within communication or the use of particular clinical micro-skills such as motivational interviewing techniques. This will primarily be dyadic, but there are some group scenarios such as group therapy, and peer support groups. These methods can also be applied to both in-person and computer-mediated settings. As well as in real patient interactions, these techniques can also be applied in therapist training, where techniques such as role-play are already widely used, and where there are also possibilities for the use of virtual patients.

Another major area of application is in the analysis of human-to-human interaction between healthcare professionals in specific settings such as surgery, and multi-disciplinary team meetings. A key feature of these scenarios is that there are distinct roles for different participants, and a number of formal practices such as the use of checklists which could impact on the ways in which CQA can be applied, perhaps in the context of a range of analysis types.

Assessment tasks are another area of application, where the output of the analysis is used directly in medical decision-making, for example in diagnosis or in the assessment of progression of disease, such as in the treatment of people with Alzheimer's.

More specifically, several applications were developed for the health industry based on the CQA. B. Renner et al. [34] proposed adaptive personalized nutrition advice systems (APNASs) that are tailored to the type and timing of personalized advice for individual needs, capacities, and receptivity in real-life food environments. They also mentioned about "participatory dialog between individuals and experts" (e.g., actual or virtual dieticians, nutritionists, and advisors) when setting goals and deriving measures of adaption. To get more information about the APNAS, the reader is kindly referred to [34].

N. Stein and K. Brooks proposed a Fully Automated Conversational Artificial Intelligence for Weight Loss [42]. They evaluated the effectiveness of a conversational AI health coach app for weight loss and behavior change, finding that users lost an average of 2.38% of their baseline weight and improved meal quality over 15 weeks of use. The results suggest that AI health coaches could be a scalable and acceptable alternative to in-person interventions for diabetes prevention and weight management, warranting further research into their potential applications in telemedicine.

Life Log Technology, Inc. provides a service, named "Calomeal advise,"[2] that provides nutrition education through dyadic dialogue with professional nutritionists by chat. The nutritionist provides advice based on users' eating habits, body condition, and their goals. From July 2023, a new feature "AI advise" is released in Calomeal Advise services. The "AI advise" analyzes the PFC (i.e., Protein, Fat, Carbohydrate) balance and priority issues of nutrient intake on a daily and weekly basis, while referring to the meal records and the most recent physical information (weight, body fat percentage, goals, etc.) of the person to be coached. Then, it automatically generates guidance comments by using ChatGPT such as personalized dietary improvement suggestions for health care each user's goals.

Mental health-care and related interventions require technologies to be developed with high standards of safety and robustness and also have a pathway towards credible evidence of clinical effectiveness. Mental health conditions such as anxiety and depression are widespread and have high personal, social, and economic costs for individuals, families, and communities. Many of those affected do not receive treatment despite available effective therapies, either through lack of access or reluctance to take up treatment options, for example, due to the stigma surrounding mental health problems. Many health systems struggle to meet the demand for mental health services, and while technologies can help to meet this need, motivation needs to be addressed in the design of technology if it is to help close the treatment gap.

[2]https://advice.calomeal.com/

Interactive Social Agents, with integrated social abilities (e.g., active listening, mimicry, gestures, emotion frameworks), increasingly form bonds with humans (Ref). Given their potential to address mental issues, these agents can be pivotal in tech-assisted therapies. Very relevant is the long-term measurement and analysis of mental health conditions and related technology-supported intervention systems using that information, as maintaining self-reporting practices over very long time periods may be challenging for clients. In this field, also many applications have already been developed and tested, including Detecting dementia via dialogue using virtual avatar [43] and SimSensei project of USC ICT [12].

The mobile socially interactive agent EmmA in the role of a vocational reintegration assistant is designed to support burn-out outpatient treatment. The system design is built upon the requirements of experts and patients. The success of such treatments is related to a patient's emotion regulation capabilities. Therefore, real-time social signal interpretation together with a computational simulation of emotion regulation influences the agent's social behavior as well as the situational selection of verbal treatment strategies [15].

Systems exist where biofeedback is technology-supported. R. Chittaro and L. Sioni [41] use virtual agents to display user stress levels. Depending on user stress detection, the socially interactive agent's emotional state and actions adjust, offering embodied feedback. Their study contrasts single vs. multi-sensor stress detection techniques, using the virtual agent's feedback to gauge the perceived accuracy of biofeedback.

T. Schneeberger and colleagues [37] introduced a virtual stress management course that uses biofeedback based on heart rate variability (HRV) and an interactive social agent as the biofeedback instructor. They assessed this system through expert interviews and a study with 71 individuals, contrasting it with conventional stress management using stress journals. Their virtual method, employing a social agent as a coach, was deemed a credible technique for teaching stress-handling strategies.

Lagos et al. [26] presented a VR-supported HRV biofeedback for golfers. Over 10 weeks, golfers and trainers used a VR golf center to hone resonance frequency breathing during play. A case study showed post-training decreases in anxiety, stress, and sensation-seeking symptoms, and increases in total HRV and sports performance.

An example of offline/asynchronous analysis that links characteristics of communication to clinical outcomes is the work presented in [10], analyzing the characteristics of successful clinical supporter messages in the context of online cognitive behavioral therapy. In this scenario, supporters review the client's progress on a scheduled basis (e.g. once per week), and provide feedback, encouragement, and guidance for the client's self-paced progress through the therapeutic intervention. In an analysis of 234,735 supporter messages, routinely collected clinical outcome data (PHQ-9 and GAD 7) is used to distinguish between more and less successful supporters, and hence identify more effective communication strategies.

Education and Training.

In the domain of education and training applications, coaching tools, especially for interviews, cater to varied dynamics. There are applications for both, dyadic and group settings. Tutoring and E-learning platforms are transforming the way we learn, complemented by applications assessing and enhancing social

skills.

The typical classroom experience is enriched through technology-supported co-teaching methods and direct teacher-student conversations or lectures. Within this area, applications for cultural understanding and language training are included. They are useful for non-native students or workers, ensuring they bridge cultural gaps. Entrance and employment examinations, including teacher evaluations, further emphasize the expansive nature of educational applications.

For the education and training field, many various applications have been developed. Formally, they can be divided into the following categories:

- **Automatic Job Interviews and the Training of Job Interviews.** Authentic Interview Prep[3] is a platform where individuals can simulate real interview scenarios by recording their responses to prompts. Given the rise of virtual interviews, this video-practice method is timely. The platform's AI analyzes key aspects of the user's presentation, from eye contact to speech patterns, pointing out areas for improvement. By reviewing these sessions, users can refine their skills, aiming for a confident and memorable performance in front of potential employers.
  TARDIS project[4] seeks to develop an immersive, scenario-driven training platform for at-risk youth aged 18-25, focusing on enhancing their social competencies. This serious game simulation employs virtual agents (VAs) that function as interviewers in mock job recruitment scenarios. These VAs are engineered to provide authentic socio-emotional interactions, serving as credible and inexhaustible conversation partners. By leveraging digital technology's unique capabilities, TARDIS creates an environment where the intensity and frequency of emotional expressions by the VAs can be adjusted. This adaptability allows for a tailored learning experience, guiding young participants through a wide spectrum of potential interview situations.
  M. Guimarães et al. [21] in their study compared interactions with an intelligent virtual character in Virtual Reality (VR) versus a traditional non-immersive platform, using a police interview scenario for Social Skills Training. Experiments were conducted in both VR and on a conventional computer screen, with data collected through presence and situated interaction questionnaires. Results revealed higher social presence of virtual characters in VR, but no significant difference in believability between the two conditions. The research suggests further investigation into measuring social presence and its impact on designing intelligent interactions for Social Skills Training in immersive environments.

- **Public Speaking.** In the CICERO project[5], the primary objective is to explore the potential enhancement of public speaking abilities through virtual training mechanisms. The methodology involves the automatic extraction of descriptions pertinent to the user's public speaking behavior using audiovisual sensors. Subsequently, an interactive virtual audience administers feedback contingent on the user's oratory performance. To understand the role of virtual characters in augmenting the educational

[3]https://innovation.ai.ets.org/products/Authentic
[4]https://cordis.europa.eu/project/id/288578
[5]https://matchollet.github.io/project/cicero/

efficacy of a public speaking training system, a modular and adaptable architecture for interactive virtual audiences was introduced. A sequence of studies was executed to evaluate the repercussions of diverse feedback techniques on training results and user experiences [11].

- **Social Skill Training.** Social skills refer to the ability to manage verbal and nonverbal behaviors during interactions with one or more individuals. People who struggle with social skills deficits find it challenging to control their own social behavior and to interpret the social behaviors of others. To address this issue, Social Skill Training (SST) is a well-established method that aims to improve individuals' social interaction abilities and reduce social stress. SST typically involves role-playing simulations of real-life situations, with the ultimate goal of helping participants become more comfortable in genuine social settings.
  As an example, TAPAS project (https://ahcweb01.naist.jp/en/projects/anr-crest-tapas/) aims to create virtual agents that can replicate the role of human SST specialists. The research also involves a detailed analysis of human social skills, breaking them down into various components, and developing specific training methods. The overarching goal is to develop effective tools and techniques to reduce social stress in everyday situations, including situations such as public speaking in educational and workplace settings.
  My Automated Conversation coacH (MACH) [22] is a system offering widespread access to social skills training, featuring a virtual agent that interprets and reacts to facial expressions, speech, and prosody. This paper highlights MACH's application in job interview training. MACH poses interview questions, mirrors specific user behaviors, and displays appropriate nonverbal responses. Post-interaction, it offers performance feedback.

- **Cross-culture Communication.** There are also many projects devoted to cross-cultural communication. As an example, the eCUTE project[6] aims to address cultural awareness challenges through innovative technology-enhanced learning methods. It will develop virtual world simulations featuring intelligent, interactive characters that model culturally specific behaviors. These scenarios, created through user-centered design, will provide immersive experiences to enhance cultural understanding and competence.

Group Collaboration.

In the area of group collaboration, creative communication, and brainstorming are paramount, fostering innovative ideas. There's a strong emphasis on mediating and enhancing communication-related tasks to ensure smooth interactions. Human resources and related technological solutions play a role, integrating coaching to optimize team dynamics. Leveraging results from CQA, it's vital to strategically determine feedback timing, recipients, and content. Lastly, conversational group recommenders are being employed to streamline collaborative decisions and discussions.

Many various types of applications have been proposed during the last decade in this field. For example, P. Lisa et al. [3] presented the AI-assisted

[6]https://cordis.europa.eu/project/id/257666

Message Rephrasing system for improving conversational qualities. The system discussed in the provided text aims to enhance conversational qualities in online interactions. It recognizes that online conversations often suffer from divisiveness and toxicity, which can negatively affect society. While efforts to improve offline conversations have been promoted, scaling these interventions to the online sphere is challenging. The text describes a large-scale experiment using artificial intelligence tools, specifically a language model, to provide real-time evidence-based recommendations to improve participants' sense of being understood in online conversations. The results show that these interventions enhance the quality of conversation, reduce political divisiveness, and improve the overall tone without altering the conversation's content or participants' policy attitudes. This research suggests promising possibilities for the use of artificial intelligence in improving the quality of online discourse and its potential impact on social media, political deliberation, and computational social science.

A. Androutsopoulou et al. [2] introduced chatbots to improve communication between government and citizens. This system aims to improve conversational qualities between the government and citizens in the public sector by utilizing AI technology, particularly chatbots. While government agencies have started adopting AI technologies based on private sector success stories, this paper emphasizes the need for extensive research to fully harness AI's potential in addressing critical public sector issues. The proposed approach leverages natural language processing, machine learning, and data mining technologies to create a new digital channel of communication between citizens and the government. It utilizes various forms of existing data, including legislation documents, structured government data, and social media data, to enable more expressive and informative interactions in everyday language. This approach is designed to handle a wider range of citizen interactions, including those with higher complexity, ambiguity, and uncertainty. The system's effectiveness has been validated through application scenarios in collaboration with Greek government agencies.

Tele-customer Services.

Within that area, methods of CQC are employed to support complaint handling, and online Q & A bots. A. Følstad et al. [13] proposed a system aimed to improve conversational qualities in chatbot interactions by addressing the issue of misinterpretations and false positives. Chatbots often struggle to accurately understand user requests, potentially leading to incorrect responses. To mitigate this problem, the study explores a strategy where the chatbot expresses uncertainty and suggests likely alternatives when its confidence in its predictions falls below a certain threshold. The research involved implementing this approach in a live chatbot for customer service and analyzing 700 chatbot dialogues before and after its implementation.

Preliminary findings suggest that introducing this solution for conversational repair can significantly reduce the occurrence of false positives in chatbot dialogues. Interestingly, expressing uncertainty and suggesting alternatives does not seem to negatively impact the overall dialogue process or the likelihood of achieving a successful outcome. These findings have theoretical and practical implications for improving chatbot interactions, and they suggest potential directions for future research in this area. Overall, the system's ability to express uncertainty and offer alternatives can enhance the quality of conversations and reduce misunderstandings.

R. Schuetzler et al. [35] proposed a system that seeks to enhance conversational qualities by improving the conversational skills of conversational agents (CAs), often referred to as chatbots. CAs are computer systems that use natural language processing to engage in conversations with humans, serving various purposes like technical support and customer service. Despite their widespread use, there has been limited research on how enhancing a CA's conversational skills affects user perceptions of the agent. The research utilizes the Social Presence Theory to explain how conversational skill influences users' perceptions of social presence and the anthropomorphism of chatbots. Through a series of studies involving 450 participants interacting with CAs of varying conversational skill levels, the research demonstrates that people perceive a more skilled CA as being more socially present and anthropomorphic compared to a less skilled CA. This study contributes to the understanding of human-computer interaction within information systems, shifting the focus from technical challenges to how users interact with CAs and how improving conversational skills can positively impact user experiences.

Computer Games.

Within that area, CQA supports multiparty conversations among non-player characters and player(s). It also supports teamwork in Massive Multiplayer Online Games (MMOG)

Many games focus on character interaction, such as role-playing games or point-and-click adventure games. However, in many commercial games, it is still common to provide players with prefabricated dialogues while free and mixed-initiative dialogs are not provided.

Not only single player games but also NPCs (Non-player characters) are adopted in a more recent trend of the game industry, MMOGs (Massive Multiplayer Online Games). The purpose of introducing such NPCs is to keep the games enjoyable and help player retention. However, an investigation on the influence of NPCs, or villagers in a greatly successful MMOG, Animal Crossing New Horizon (ACNH) reported the limited dialogue capability, they do not really contribute to the game in this aspect [Hsieh 2021].

Many multiplayer online games offer their players in-game text or voice and video chat. Depending on game and genre, chat may be an essential part of the game mechanics, as it allows groups of players to coordinate their actions, whereas in other games the chat is used mainly to allow players to socialize during gameplay. There are also chatting-apps, that feature some built-in games.

S. Yildirim et al. [44] proposed a system aimed to improve conversational qualities in spoken dialog systems by automatically recognizing the user's communicative style, including affective aspects such as frustration, politeness, and neutrality. The goal is to make interactions with dialog systems richer and more natural by understanding not only what is said but also how it is communicated. The study uses various information sources, including acoustic, lexical, and contextual features, to detect these communicative styles in children's speech during spontaneous dialogues with computer characters. The research involved a corpus of 103 children aged 7-14 playing a voice-activated computer game. The experimental results indicate that lexical information is particularly effective in detecting politeness, while context and acoustic features are more suitable for detecting frustration. Combining these different information sources leads to significantly improved classification results. Additionally, the study found that classification performance varies with age and gender, with higher accuracy in

detecting politeness among females and 10-11-year-olds compared to males and other age groups. This research contributes to enhancing conversational qualities in dialog systems by better understanding and responding to the emotional and communicative nuances in users' speech.

4.4.2 Future Work

Much work remains to be done to look at the integration of CQA functionality into real-world applications, looking at how the results of CQA should be presented to different stakeholders, and dealing with issues of consent. We identified a few domains which are aligned with the SOTA above, elaborated on some future commercial applications in each of those domains, and touch on some technical, ethical and legal issues surrounding those domains.

4.5 Ethics and Societal Impact

There are a variety of ethical and societal implications that can arise from the utilization of CQs.

On the negative side, CQ has the potential to influence a user without their conscious realization, and CQA can extract private information from them without their awareness. There's the danger of persuasive conversations and manipulation of conversation qualities, where the user may be unduly influenced without their knowledge. Such actions raise significant concerns about personal privacy, especially when there's a risk of privacy violations through conversations and self-disclosures. Moreover, there's potential for psychological harm, where users could be exposed to negative emotional stimuli, shamed, or asked uncomfortable questions during interactions. The current data collection practices show inherent biases against certain demographic groups, such as specific ages, genders, or ethnicities. When these biases become a part of models, they can perpetuate and amplify pre-existing societal biases. Over-reliance or dependency on these technologies can lead to addiction, and there's a concern that technology might serve as a substitute for human interactions, leading to societal isolation. Accessibility and fairness are also in question; the technologies might be limited to only affluent users, excluding those who cannot afford them. Lastly, there's a potential risk of these technologies being repurposed for unethical uses.

However, there are also positive outcomes. The advancements in CQ and CQA can lead to the dissemination of knowledge, enhancing public and professional awareness about the importance of nonverbal cues in communication. These technologies can offer an effortless entry point for interventions, potentially leading to positive impacts on users' well-being. The early detection of biases in critical systems is another significant benefit, alongside the potential for increased transparency concerning the implications of CQA. Users can also derive enjoyment and enhanced experiences from CQ technologies.

There are also neutral or yet-to-be-determined implications and questions associated with these advancements. For instance, how will conversational dynamics evolve if people can't easily hide their genuine feelings or thoughts? The potential and controversial uses of CQA in assessing individuals' trustworthiness, matchmaking, group formation, or employment decisions raise ethical concerns. Questions also arise about who has the right to assess, access, and use

CQA information, and who bears the liability if CQA provides incorrect data or causes harm. Lastly, the impact of CQA on various spheres, such as interaction, task, life of use, and society, needs thorough exploration and understanding.

4.5.1 How to Deal with Ethical, Legal and Societal Issues

Dealing with ethical, legal, and societal issues requires a multifaceted approach. It's crucial to comply with both national and international standards, seeking approval from relevant boards that oversee ethical, GDPR, legal, and safety concerns. Alongside these standards, adhering to established best practices and recommendations is essential.

Specific features related to CQ should be given special attention. This includes ensuring there's no discrimination and that data and models used for CQ and CQA are balanced to maintain fairness. Care must be taken when working with vulnerable participants, and they should be informed about any CQ manipulations and the methods of CQA.

Moreover, it's important to think about the long-term implications of CQ and CQA, ensuring that there's sustained oversight and responsibility, especially as concerns arise about potential long-term challenges, like addiction to technology. Efforts should be made to disseminate knowledge about CQA, raising awareness about the technology. This can be achieved through responsible design guidelines, integrating CQA into educational curricula, and providing appropriate training. A vital part of this awareness is enhancing the transparency and explainability of systems that utilize CQA. Individuals should clearly understand how their data is used, where it's stored, and by whom. There should also be explicit communication about the potential risks of inappropriate applications, ensuring the public is informed about both the negative and positive impacts of CQ and CQA.

Differentiating between the challenges faced during research and the design of applications and interventions is important. For research, there's a need to consider the long-term implications, the potential effects of new findings, and the insights participants may gain about themselves. When designing applications and interventions, ethical considerations should be integrated from the onset, embodying an "ethical by design" philosophy. Emphasis should also be placed on ensuring opt-in and opt-out, allowing users to opt-in or opt-out, and potentially employing blockchain technologies to ensure that data control remains with the individual from whom the data originates.

Lastly, the social responsibility associated with CQ research should never be overlooked. Researchers must be aware of the potential societal applications of their work, and understand the kinds of tools and services that could arise from it. They should also consider how to apply their findings responsibly in practical applications, like using personality tests to assess specific skills or traits, ensuring the trustworthiness and reliability of the results in real-world contexts.

List of Participants

- Shogo Okada, Japan Advanced Institute of Science and Technology
- Yukiko Nakano, Seikei University
- Elisabeth André, Augsburg University
- Wolfgang Minker, Ulm University
- Yuki Matsuda, Nara Institute of Science and Technology (NAIST)
- Denis Dresvyanskiy, Ulm University
- Matthias Kraus, Augsburg University
- Albert Ali Salah, University of Utrecht
- Catherine Pelachaud, Institut Systèmes Intelligents et de Robotique (ISIR)
- Shirou Kumano, Nippon Telegraph and Telephone Corporation: NTT
- Sakriani Sakti, Japan Advanced Institute of Science and Technology (JAIST)
- Martin Baumann, Ulm University
- Jean-Claude Martin, CNRS-LIMSI
- Huang Hung-Hsuan, Fukuchiyama University
- Heysem Kaya, Utrecht University
- Thomas Rist, Augsburg University of Applied Sciences
- Patrick Gebhard, DFKI
- Yugo Nakamura, Kyushu University
- Gavin Doherty, Trinity College, Dublin
- Abhinav Dhall, IIT Ropar
- Leong Chee Wee, Educational Testing Service (ETS)
- David Traum, University of Southern California
- Kristiina Jokinen, AIST
- Graham Wilkock, CDM Interact Oy.
- Takayuki Nozawa, University of Toyama
- Quan Jingyu, Tokyo Institute of Technology
- Marius Funk, University of Augsburg
- Koichiro Yoshino, RIKEN

References

[1] Alameda-Pineda, X., Yan, Y., Ricci, E., Lanz, O., Sebe, N.: Analyzing free-standing conversational groups: A multimodal approach. In: Proceedings of the 23rd ACM International Conference on Multimedia. p. 5–14. MM '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2733373.2806238

[2] Androutsopoulou, A., Karacapilidis, N., Loukis, E., Charalabidis, Y.: Transforming the communication between citizens and government through ai-guided chatbots. Government Information Quarterly **36**(2), 358–367 (2019). https://doi.org/https://doi.org/10.1016/j.giq.2018.10.001

[3] Argyle, L.P., Busby, E., Gubler, J.R., Bail, C.A., Howe, T., Rytting, C., Wingate, D.: Ai chat assistants can improve conversations about divisive topics. ArXiv **abs/2302.07268** (2023)

[4] Badura, K.L., Galvin, B.M., Lee, M.Y.: Leadership emergence: An integrative review. Journal of Applied Psychology **107**(11), 2069 (2022)

[5] Bartneck, C.: Godspeed Questionnaire Series: Translations and Usage, pp. 1–35. Springer International Publishing, Cham (2023)

[6] Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J.: An inventory for measuring depression. Archives of general psychiatry **4**(6), 561–571 (1961)

[7] Benlamine, M.S., Chaouachi, M., Villata, S., Cabrio, E., Frasson, C., Gandon, F.L.: Emotions in argumentation: an empirical evaluation. In: International Joint Conference on Artificial Intelligence (2015)

[8] Bhattacharya, I., Foley, M., Zhang, N., Zhang, T., Ku, C., Mine, C., Ji, H., Riedl, C., Welles, B.F., Radke, R.J.: A multimodal-sensor-enabled room for unobtrusive group meeting analysis. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. p. 347–355. ICMI '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3242969.3243022

[9] de Buisonjé, D.R., Ritter, S.M., de Bruin, S., ter Horst, J.M.L., Meeldijk, A.: Facilitating creative idea selection: The combined effects of self-affirmation, promotion focus and positive affect. Creativity Research Journal **29**(2), 174–181 (2017). https://doi.org/10.1080/10400419.2017.1303308

[10] Chikersal, P., Belgrave, D., Doherty, G., Enrique, A., Palacios, J.E., Richards, D., Thieme, A.: Understanding client support strategies to improve clinical outcomes in an online mental health intervention. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. p. 1–16. CHI '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3313831.3376341

[11] Chollet, M., Marsella, S., Scherer, S.: Training public speaking with virtual social interactions: effectiveness of real-time feedback and delayed feedback. Journal on Multimodal User Interfaces **16**(1), 17–29 (2022)

[12] DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., Morency, L.P.: Simsensei kiosk: a virtual human interviewer for healthcare decision support. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems. p. 1061–1068. AAMAS '14, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2014)

[13] Følstad, A., Taylor, C.: Conversational repair in chatbots for customer service: The effect of expressing uncertainty and suggesting alternatives. In: Følstad, A., Araujo, T., Papadopoulos, S., Law, E.L.C., Granmo, O.C., Luger, E., Brandtzaeg, P.B. (eds.) Chatbot Research and Design. pp. 201–214. Springer International Publishing, Cham (2020)

[14] Garfield, J.L., Peterson, C.C., Perry, T.: Social cognition, language acquisition and the development of the theory of mind. Mind & Language **16**(5), 494–541 (2001). https://doi.org/https://doi.org/10.1111/1468-0017.00180

[15] Gebhard, P., Schneeberger, T., Dietz, M., André, E., Bajwa, N.u.H.: Designing a mobile social and vocational reintegration assistant for burnout outpatient treatment. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. p. 13–15. IVA '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3308532.3329460

[16] Ghosh, A., Chakraborty, D., Prasad, D., Saha, M., Saha, S.: Can we recognize multiple human group activities using ultrasonic sensors? In: 2018 10th International Conference on Communication Systems & Networks (COMSNETS). pp. 557–560 (2018). https://doi.org/10.1109/COMSNETS.2018.8328272

[17] Ghosh, S., Dhall, A., Hayat, M., Knibbe, J., Ji, Q.: Automatic gaze analysis: A survey of deep learning based approaches. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**, 61–84 (2021)

[18] Gordon, D., Hanne, J.H., Berchtold, M., Miyaki, T., Beigl, M.: Recognizing group activities using wearable sensors. In: Puiatti, A., Gu, T. (eds.) Mobile and Ubiquitous Systems: Computing, Networking, and Services. pp. 350–361. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)

[19] Gravano, A., Hirschberg, J.: A corpus-based study of interruptions in spoken dialogue. In: Interspeech (2012)

[20] Gross, K., Brewer, P.R.: Sore losers: News frames, policy debates, and emotions. Harvard International Journal of Press/Politics **12**(1), 122–133 (2007). https://doi.org/10.1177/1081180X06297231

[21] Guimarães, M., Prada, R., Santos, P.A., Dias, J.a., Jhala, A., Mascarenhas, S.: The impact of virtual reality in the social presence of a virtual agent. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents. IVA '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3383652.3423879

[22] Hoque, M.E., Courgeon, M., Martin, J.C., Mutlu, B., Picard, R.W.: Mach: my automated conversation coach. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing. p. 697–706. UbiComp '13, Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2493432.2493502

[23] Kroenke, K., Spitzer, R.L., Williams, J.B.: The phq-9: validity of a brief depression severity measure. Journal of general internal medicine **16**(9), 606–613 (2001)

[24] Kroenke, K., Strine, T.W., Spitzer, R.L., Williams, J.B., Berry, J.T., Mokdad, A.H.: The phq-8 as a measure of current depression in the general population. Journal of Affective Disorders **114**(1), 163–173 (2009). https://doi.org/https://doi.org/10.1016/j.jad.2008.06.026

[25] Krych-Appelbaum, M., Law, J.B., Jones, D., Barnacz, A., Johnson, A., Keenan, J.P.: "I think I know what you mean": The role of theory of mind in collaborative communication. Interaction Studies **8**(2), 267–280 (2007). https://doi.org/https://doi.org/10.1075/is.8.2.05kry

[26] Lagos, L., Vaschillo, E., Vaschillo, B., Lehrer, P., Bates, M., Pandina, R.: Virtual reality–assisted heart rate variability biofeedback as a strategy to improve golf performance: a case study. Biofeedback **39**(1), 15–20 (2011)

[27] Mitsunaga, N., Smith, C., Kanda, T., Ishiguro, H., Hagita, N.: Adapting robot behavior for human–robot interaction. IEEE Transactions on Robotics **24**(4), 911–916 (2008). https://doi.org/10.1109/TRO.2008.926867

[28] Moore, G.C., Benbasat, I.: Development of an instrument to measure the perceptions of adopting an information technology innovation. Inf. Syst. Res. **2**, 192–222 (1991)

[29] Morgan, D.L., Hoffman, K.: A system for coding the interaction in focus groups and dyadic interviews. The Qualitative Report **23**(3), 519–531 (2018)

[30] Morgeson, F.P., DeRue, D.S., Karam, E.P.: Leadership in teams: A functional approach to understanding leadership structures and processes. Journal of Management **36**(1), 5–39 (2010). https://doi.org/10.1177/0149206309347376

[31] Neff, J.J., Fulk, J., Yuan, Y.C.: Not in the mood? affective state and transactive communication. Journal of Communication **64**(5), 785–805 (2014). https://doi.org/https://doi.org/10.1111/jcom.12109

[32] Padilha, E., Carletta, J.: A simulation of small group discussion (2002), 6th Workshop on the Semantics and Pragmatics of Dialogue (EDILOG 2002) ; Conference date: 04-09-2002 Through 06-09-2002

[33] Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.Y., Wang, W.Y.: On the risk of misinformation pollution with large language models. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023)

[34] Renner, B., Buyken, A.E., Gedrich, K., Lorkowski, S., Watzl, B., Linseisen, J., Daniel, H., Conrad, J., Ferrario, P.G., Holzapfel, C., Leitzmann, M., Richter, M., Simon, M.C., Sina, C., Wirsam, J.: Perspective: A conceptual framework for adaptive personalized nutrition advice systems (apnass). Advances in Nutrition **14**(5), 983–994 (2023). https://doi.org/https://doi.org/10.1016/j.advnut.2023.06.009

[35] Ryan M. Schuetzler, G.M.G., Giboney, J.S.: The impact of chatbot conversational skill on engagement and perceived humanness. Journal of Management Information Systems **37**(3), 875–900 (2020). https://doi.org/10.1080/07421222.2020.1790204

[36] Sanchez-Cortes, D., Aran, O., Jayagopi, D.B., Schmid Mast, M., Gatica-Perez, D.: Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. Journal on Multimodal User Interfaces **7**, 39–53 (2013)

[37] Schneeberger, T., Sauerwein, N., Anglet, M.S., Gebhard, P.: Stress management training using biofeedback guided by social agents. In: Proceedings of the 26th International Conference on Intelligent User Interfaces. p. 564–574. IUI '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3397481.3450683

[38] Sebo, S., Stoll, B., Scassellati, B., Jung, M.F.: Robots in groups and teams: A literature review. Proc. ACM Hum.-Comput. Interact. **4**(CSCW2) (oct 2020). https://doi.org/10.1145/3415247

[39] Shiffman, S., Stone, A.A., Hufford, M.R.: Ecological momentary assessment. Annu. Rev. Clin. Psychol. **4**(1), 1–32 (2008)

[40] Shriberg, E., Stolcke, A., Baron, D.: Observations on overlap: findings and implications for automatic processing of multi-party conversation. In: Interspeech (2001)

[41] Sioni, R., Chittaro, L.: Stress detection using physiological sensors. Computer **48**(10), 26–33 (2015). https://doi.org/10.1109/MC.2015.316

[42] Stein, N., Brooks, K., et al.: A fully automated conversational artificial intelligence for weight loss: longitudinal observational study among overweight and obese adults. JMIR diabetes **2**(2), e8590 (2017)

[43] Tanaka, H., Adachi, H., Ukita, N., Ikeda, M., Kazui, H., Kudo, T., Nakamura, S.: Detecting dementia through interactive computer avatars. IEEE Journal of Translational Engineering in Health and Medicine **5**, 1–11 (2017). https://doi.org/10.1109/JTEHM.2017.2752152

[44] Yildirim, S., Narayanan, S., Potamianos, A.: Detecting emotional state of a child in a conversational computer game. Computer Speech & Language **25**(1), 29–44 (2011). https://doi.org/https://doi.org/10.1016/j.csl.2009.12.004, affective Speech in Real-Life Interactions

[45] Zigurs, I., Kozar, K.A.: An exploratory study of roles in computer-supported groups. MIS Quarterly **18**(3), 277–297 (1994)

[46] van Zoelen, E.M., van den Bosch, K., Rauterberg, M., Barakova, E., Neerincx, M.: Identifying interaction patterns of tangible co-adaptations in human-robot team behaviors. Frontiers in Psychology **12** (2021). https://doi.org/10.3389/fpsyg.2021.645545