UMEÅ UNIVERSITY

Breakdown Situations in Dialogues Between Humans and Socially Intelligent Agents

Maitreyee Tewari

DOCTORAL THESIS, DECEMBER 2023
DEPARTMENT OF COMPUTING SCIENCE
UMEÅ UNIVERSITY
SWEDEN

Department of Computing Science
Umeå University
SE-901 87 Umeå, Sweden

maitreyee.maitreyee@umu.se

ISBN 978-91-8070-164-8 (print)
ISBN 978-91-8070-165-5 (pdf)
ISSN 0348-0542
UMINF 23.07

Front cover by Maitreyee Tewari and Ida Åberg
Printed by CityPrint i Norr AB, 2023

Abstract

Dialogues between humans are complex due to the challenges in predicting how they will unfold as people may want to achieve different purposes. For instance, to act together, they co-create a common goal; to learn, they co-create knowledge; to build relationships, they share emotions and beliefs. Apart from different purposes, people may want to achieve multiple purposes in a dialogue, introducing a movement between goals. Such actions cause problems in understanding and conflicts among the participants. Activity Theory denotes such situations as breakdown situations, which also occur when people have dialogues with software agents driven by Artificial Intelligence (AI). This thesis falls within the domain of human-centred AI, focusing on software agents able to collaborate and support people to achieve their goals. We call these software agents socially intelligent agents.

This thesis has two aims: (1) to develop an increased understanding of breakdown situations in dialogues between humans and socially intelligent agents and (2) to develop computational frameworks based on the developed understanding to manage breakdown situations, which could be embedded in an agent's cognitive architecture. The theoretical frameworks from social sciences, particularly Activity Theory, were applied to address the aims. They provided an alternate perspective that considers breakdown situations as opportunities to learn something new rather than the traditional view of them being errors or failures.

The main contributions addressing the first aim were theory-driven analysis and empirical findings that provided increased knowledge of breakdown situations, resulting in design implications and future agendas guiding the subsequent research. The results informed the three strategies to manage breakdown situations by aligning, partially aligning or not aligning with human's intentions. We found that participants considered partial alignment as a sufficient level of agreement for potential collaboration, which would be interesting to verify in future studies. To address the second aim, two novel computational frameworks were provided. These frameworks were based on linguistics and social sciences theories, allowing an agent to interpret the dialogue's syntax, semantics, and social aspects, facilitating a deeper understanding of dialogues. Finally, a novel computational framework was developed to reason about conflicts and be able to plan by adopting the strategy of aligning with the human's intentions.

We conceptualised a cognitive architecture based on our research findings. The cognitive architecture embeds mechanisms for socially intelligent agents to manage breakdown situations in dialogues with humans.

Sammanfattning

Dialoger mellan människor är komplexa eftersom det är svårt att förutse hur de kommer att utveckla sig när människor ofta har olika syften med dialogen. Exempelvis, för att utföra något tillsammans samskapar man ett gemensamt mål; för att lära, samskapar man kunskap; för att bygga relationer delar man upplevelser, vad man känner och vill. Dessutom kan man vilja uppnå flera mål i en dialog, så att fokus flyttas mellan flera syften. Sådana företeelser kan orsaka problem med att förstå och konflikter mellan deltagarna. Verskamhetsteorin definierar sådana situationer sammanbrottssituationer. Sådana situationer uppstår också när människor kommunicerar med mjukvaruagenter byggda med teknik från området artificiell intelligens (AI). Denna avhandling faller inom området människo-centrerad AI som fokuserar på mjukvaruagenter som kan samarbeta och stödja människor att uppnå sina mål. Vi kallar sådana mjukvaruagenter socialt intelligenta agenter.

Denna avhandling har två syften: (1) att öka förståelsen av sammanbrottssituationer i dialoger mellan människor och socialt intelligenta agenter; och (2) att utveckla beräkningsmetoder baserade på insikterna för att identifiera och hantera sammanbrottssituationer, vilka skulle kunna integreras i en agents kognitiva arkitektur.

Teoretiska ramverk, särskilt verksamhetsteori, används, vilka ger ett alternativt perspektiv jämfört med det traditionella perspektivet att se sammanbrottssituationer som orsakade av felaktigheter hos mjukvaran. Istället ses sådana situationer som tillfällen att skapa ny kunskap och att utvecklas, inte bara för människan utan också för mjukvaruagenten.

Resultaten relaterade till det första syftet innefattar en teoribaserad analys och empiriska resultat som gav en ökad förståelse av sammanbrottssituationer, vilket resulterade i designimplikationer och en forskningsagenda vilka i sin tur informerade den fortsatta forskningen. Denna innefattar tre strategier för att hantera sammanbrottsituationer: att följa, delvis följa, eller inte följa personens avsikter beroende på situationen. Vi fann att forskningspersoner uppfattade delvis efterföljande som tillräcklig nivå av samarbete, vilket är intressant att verifiera i framtida studier.

För att möta det andra syftet utvecklades två ramverk för beräkning baserade på lingvistiska och samhällsvetenskapliga teorier. Dessa ramverk kan agenten använda för att analysera dialogens syntax, semantik, och sociala aspekter, i syfte att skapa en djupare förståelse för situationen. Dessutom utvecklades en beräkningsmetod för att agenten ska kunna resonera kring konflikter och planera utifrån strategin att följa människans avsikter.

Denna avhandling bidrar även med en konceptuell modell av en kognitiv arkitektur baserad på forskningen som bygger in mekanismer för att en socialt intelligent agent ska kunna identifiera och hantera sammanbrottssituationer i dialoger med människan.

Acknowledgment

I would like to start by thanking my supervisor, Helena Lindgren, for her guidance and support during my studies. I would also like to acknowledge my co-supervisor, Kai-Florian Richter and reference person, Eddie Wadbro. Their guidance enabled me to be a better researcher.

I am grateful to all the researchers– Helena, Kai-Florian, Michele, Esteban, Thomas, and Suna with whom I had the opportunity to collaborate during my studies. The collaboration played an important role in shaping my research.

I would like to acknowledge my husband, Michele, for his support and insights during our many research-related discussions. I would also like to acknowledge our families for their affection, support, and confidence in my abilities.

I am thankful to my old-time friends– Aishwarya, Suraj, Varun, Nagendra, and Vinu, for their inspiration and encouragement. A great thanks to those whom I met during my PhD studies– Luis, Cecília, Eunil Seo, Anouk, Antonio, Aleksandar, Monika, Abel, Peter, Ola, Polina, and Esteban for their affection and kindness.

Finally, I want to acknowledge the efforts by the management at the Institutionen för datavetenskap to ensure a functional environment to conduct my studies. I am also grateful to Patrik Eklund, Erik Elmroth, Late. Lena Kallin Westin and Frank Dignum for their guidance contributing to my professional development.

Funding and Provision

- Work done in I, II, V, and VI was partially funded by the Humane-AI-Net excellence network funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 952026.
- Work done in papers III and IV was funded by the Socrates Project, a European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721619.

Preface

Papers included in this thesis.

Paper I — **Maitreyee Tewari** and Helena Lindgren. Younger and Older Adults' Perceptions on Role, Behavior, Goal and Recovery Strategies for Managing Breakdown Situations in Human-Robot Dialogues. *In Proceedings of the 9th International Conference on Human-Agent Interaction, ACM Digital Library, Pp. 433-437, 2021.*

Paper II — **Maitreyee Tewari** and Helena Lindgren. Expecting, Understanding, Relating and Interacting - Older, Middle-aged, and Younger Adults' Perspectives on Breakdown Situations in Human-Robot Dialogues. *Frontiers in Robotics and AI, Frontiers Media SA, Volume 9: 956709, 2022.*

Paper III — **Maitreyee Tewari**, Suna Bensch, Thomas Hellström and Kai-Florian Richter. Modelling Grice's Maxim of Quantity as Informativeness for Short Text. *In ICLLL 2020: The 10th International Conference in Languages, Literature and Linguistics, Pp. 1-7. 2020.*

Paper IV — **Maitreyee Tewari** and Michele Persiani. Variational Autoencoding Dialogue Sub-Structures Using a Novel Hierarchical Annotation Schema. *In Proceedings of the 6th IEEE Congress on Information Science and Technology (CiSt), IEEE, Pp. 334-341, 2020.*

Paper V — Esteban Guerrero, **Maitreyee Tewari,** Helena Lindgren and Panu Kalmi. Forming *We-intentions* under Breakdown Situations in Human-Robot Interactions. *Computer Methods and Programs in Biomedicine, Elsevier, Volume 242: 107817, 2023.*

Paper VI — **Maitreyee Tewari** and Michele Persiani. Towards We-intentional Human-Robot Interaction using Theory of Mind and Hierarchical Task Network. *In Proceedings of the 5th International Conference on Computer-Human Interaction Research and Applications - Humanoid, SciTePress, Pp. 291-299, 2021.*

Other articles and technical reports not part of the thesis that were produced during the PhD studies:

- Maitreyee Tewari. Formalization of Dialogues from Movie Corpus using DAMSL Annotation Scheme as Cooperating Distributed Grammar Systems. *Report / UMINF 21.07. 2021.*

- Michele Persiani and Maitreyee Tewari. Mediating Joint Intention with a Dialogue Management System. *In 1st International Workshop on New Foundations for Human-Centered AI, Pp. 79-82, 2020.*

- Maitreyee Tewari. Beyond Adjacency Pairs: Hierarchical Clustering of Long Sequences for Human-Machine Dialogues. *In Proceedings of the 1st Workshop on Computational Approaches to Discourse, ACM Digital Library, Pp. 11-19, 2020.*

- Maitreyee Tewari, Monika Jingar and Suna Bensch. A Hybrid Model to Classify Sudden Topic Change, Misunderstanding and Non-understanding in Human Chat-bot Interaction. *Preprint manuscript on Diva at webpage* `http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-174674`, *2020.*

- Maitreyee Tewari and Suna Bensch. Natural Language Communication with Social Robots for Assisted Living. *In Robots for Assisted Living- 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, Pp. 1-4, 2018.*

The following is a summary of the contributions Tewari made to each paper included in this thesis.

- In papers I, II, and III, Tewari led the research from conceptualisation of the problem, implementation of the computational framework in Paper III to designing, setting up, conducting the studies, analysing the results, and drafting the initial version of the papers. Supervisors guided throughout the research process and helped in the final drafting of the papers.
- In Paper IV, Tewari conceptualised the problem, annotated the dataset, performed the experiments, analysed the results and prepared the initial draft.
- In Paper V, Tewari conceptualised the problem based on previous studies conducted in papers I and II. Tewari's main contribution was to evaluate the developed strategies for managing breakdown situations in dialogues. Tewari performed the analysis and reporting of the results. Tewari contributed to authoring the paper and maturing it for publication.
- In Paper VI, Tewari conceptualised the problem, contributed to designing the computational framework and co-authored the paper.

Contents

Chapter 1

Introduction

> "(...) If a machine is expected to be infallible, it cannot also be intelligent. (...)"
>
> *Alan Turing*

Communicating using dialogues is a crucial means for humans to achieve goals together, express feelings, resolve conflicts, and build understanding about a topic being discussed or a task to be completed.

When software agents driven by Artificial Intelligence (AI) engage in dialogues with humans, they must behave socially appropriately and manage the dialogue intelligently [26, 55]. In this thesis, we call such agent *socially intelligent agent*. Although this thesis focuses on dialogues, we include embodiment as body language (physical or virtual) typically supports dialogues.

Enabling socially appropriate and intelligent behaviour requires agents to create an understanding of dialogues similar to humans' understanding. However, creating such an understanding is a challenge in AI research. This challenge is due to the difficulty in predicting the unfolding of a dialogue, introduced by situations when intentions conflict or when there are problems due to insufficient or inconsistent knowledge about, for example, the facts, norms, and emotions of the other actors [38, 55].

So far, the conflicts and problems in understanding have been addressed mainly from the engineering perspective to reduce errors [39, 51, 54, 70, 77]. We are interested in exploring how a human-centred approach can improve our understanding of

conflicts and problems for socially intelligent agents who would potentially engage in dialogues with humans.

A first step to exploring this human-centred understanding includes understanding conflicts of intentions and inconsistent or lack of knowledge about emotions, procedures, rules, intentions, and norms. We group and refer to these situations as *breakdown situations*.

The breakdown situations are explored by adopting an interdisciplinary perspective on purposeful human activity in general and purposeful dialogue activities involving a human and a socially intelligent agent [22, 48, 95]. We assume that breakdown situations are intrinsic to dialogues socially intelligent agents would potentially have with humans, similar to dialogues between humans [38, 56]. Furthermore, these situations are seen as *opportunities to learn* something new and contribute to the development of the human counterpart, the socially intelligent agent, and the joint activity [7].

As a starting point, this research explores what *understanding* entails in dialogues between humans. Linell [50] defines an *understanding* to be an outcome that establishes the relation between the *utterance*, the speaker (A), the receiver (B), and the *world*. An utterance is not only for B to understand what A intends, feels, or knows but also for A to understand better what they intend, feel, or know. Understanding is activity-specific and an act of fitting the utterance and connecting it to a *context*. Where the context provides a concrete setting, framework, and knowledge structure that gives meaning to utterances and allows a participant to create a relevant interpretation of the utterance [21].

Human understanding of dialogues involving other humans has been described as partial and incomplete, often leading to breakdown situations. Humans typically embed the creation of a mutual agreement in dialogues and have strategies to recover from breakdown situations to manage understanding as the dialogue unfolds [38, 50]. Previous research involving dialogues between humans and socially intelligent agents has focused on creating mechanisms to understand and manage breakdown situations related to either speech [29, 69, 71], factual knowledge (such as names, participants, and objects), or intentions [1, 28, 57, 77, 78]. Instead of focusing on a specific aspect of either intentions, knowledge, or speech, this work is interested in exploring the diverse aspects required and embedded in creating an *understanding* based on a mutual agreement. The outcome is expected to specify the requirements for developing mechanisms that agents can use to create a *deeper understanding* to manage dialogues and related breakdown situations.

Researchers have explored how the erroneous or faulty behaviour of socially intelligent agents performing tasks and participating in task-specific dialogues with humans are perceived [58, 63, 67]. However, what has been less explored is

people's perception of and expectations from socially intelligent agents engaging in dialogues about daily activities and health, which embed breakdown situations and their management. Therefore, we are interested in exploring how people *perceive* and what they *expect* from a socially intelligent agent. The outcome is expected to inform the design of how socially intelligent agents could and should *behave* when faced with breakdown situations in dialogues with humans.

1.1 Objectives and Research Questions

The *objective* of the thesis is two-folded: (1) develop an increased understanding of breakdown situations in human-agent dialogues, which could inform the design of socially intelligent agents; and (2) develop computational frameworks to recognise and manage these situations, to deepen the socially intelligent agent's understanding of breakdowns.

The research in this thesis addresses the following research questions:

RQ1 "How do people **perceive** breakdown situations and their management in dialogues between a human and a socially intelligent agent?"

RQ2 "What do people **expect** from a socially intelligent agent when breakdown situations occur in dialogues with humans?"

RQ3 "How can the socially intelligent agent **recognise** a breakdown situation has occurred during a dialogue with the human?"

RQ4 "What strategies and techniques can the socially intelligent agent employ to **manage** breakdown situations in a dialogue with the human?"

1.2 Research Contributions

The following summarises this thesis's research contributions towards addressing the research questions.

Paper I [87] contributes with the following:

- An increased knowledge of *roles and relationships, understanding and emotional connection* and *behaviour and adaptive manner* of socially intelligent agents.

– A set of design implications embedding mechanisms that allow *co-construction of knowledge and understanding, management and prevention of breakdown situations, transparent decision-making,* and *self-reflection* in the development of the socially intelligent agents.

Paper II [86] contributes with the following:

– Activity Theory-based methodology to study breakdown situations.
– Increased knowledge along an over-arching theme about participant *expectations* consisting of three aspects– *understanding, interacting, relating*. The expectations were categorised into what the socially intelligent agent needs in order to develop the *understanding* about the ongoing dialogue, how it could *interact naturally* and how it could *relate* with the participant. The themes were refined and segmented into factors composing the three aspects.
– Design implications and a research agenda guiding the future research in developing socially intelligent agents to collaborate and conduct dialogues with humans while managing breakdown situations.

Paper III [84] contributes with:

– A novel computational framework that measures *informativeness* and *syntactic cohesion* in dialogues.
– The evaluation of the framework showed alignment between the scores produced by the framework and those of the participants. This framework could be applied to predict and prevent breakdown situations in the future.

Paper IV [89] contributes with:

– A novel *hierarchical annotation schema* for capturing sequences of dialogues longer than two turns. The annotation schema was used to annotate a dataset of dialogues with aspects relating to syntax, semantics, and social norms.
– A dataset composed of chit-chat and task-driven dialogues between two humans and between humans and machines was annotated with the three aspects.
– The evaluation of the annotation schema showed that generating sequences up to five turns long was possible. The proposed annotation schema could be applied to recognise breakdown situations occurring

at syntax, semantics, and social levels of understanding. The generative model could also generate dialogue turns that agents can apply to manage breakdown situations.

Paper V [35] contributes with the following:

- A novel computational framework with reasoning mechanisms to recognise and manage conflict of intentions.
- Definitions of three strategies to manage conflict of intentions by *aligning, partially aligning,* and *not aligning* to other participants' intentions.
- Preliminary evaluations showed participants could recognise when the human and the embodied agent aligned and performed a joint activity or did not align and rejected the proposed joint activity. An interesting observation was participants considered partial alignment as a *good enough agreement* for potential joint activity.

Paper VI [88] contributes with:

- A novel computational framework to reason and plan without needing to manage the conflict of intention. The agent reasons the conflict as humans aiming for their private goals (I-mode We-intention) rather than an agreed-upon group goal (We-mode We-intention). As a strategy to manage conflict of intentions, the agent always aligns with human intentions. When the agent reasons the human is operating in We-mode We-intention, it creates a plan embedding human expectations. Instead, the agent makes an optimal plan when an I-mode We-intention is detected.

1.3 Delimitation

This thesis is presented at a time when learning-based methods for dialogue management, such as chatbots, have a grip on society's attention. Such dominance of the learning-based methods through Large Language Models (LLMs) [100] for chatbots and dialogue management systems could easily lead readers to assume this thesis follows a similar research line. Therefore, this section is dedicated to summarising what this thesis is about and what it is not.

- The thesis focuses on the problem of breakdown situations in purposeful human-agent dialogues.

- The explorations are done from an interdisciplinary perspective. Involving theories from social sciences, psychology, philosophy and computer science.
- This thesis's primary goal is to develop agents capable of collaborating with humans, learning and supporting them, and making decisions that are transparent to humans, following the human-centric AI perspective [61]. This differs from LLMs that do not embed mechanisms to allow transparency and purposeful reasoning over time [100].
- This thesis is not primarily about building dialogue management systems. Instead, about creating a deeper understanding of what aspects need to be embedded to facilitate natural dialogues, including the management of breakdown situations. Furthermore, computational frameworks driven by human-centric recommendations for managing dialogues were explored.
- The studies were conducted with socially intelligent embodied agents in home-care scenarios, which sets the stage for our user studies and delimits the use case.

1.4 Thesis Overview

The thesis is organised as follows: Chapter 2 introduces relevant theories underlying this research. In Chapter 3, we identify various reasons for breakdown situations, perform theory-driven analysis of those reasons, build dialogue scenarios embedding them, and set up studies to understand people's opinions on those dialogue scenarios. Chapter 4 focuses on building computational mechanisms for recognising breakdown situations relating to syntax, semantics, and social aspects of dialogues. Chapter 5 focuses on and describes methods to recognise and manage breakdown situations due to conflict of intentions. Chapter 6 summarises the research contributions that each paper included in this thesis makes towards addressing the research questions. The contributions are further extended and feed into conceptualising a cognitive architecture. We conclude this thesis by discussing the contributions, ethical concerns, limitations, and future work in Chapter 7.

Chapter 2

Background on Theoretical Frameworks

Human-centric AI focuses on developing agents to collaborate and support humans in achieving their goals effectively [61]. To develop embodied agents following the human-centric AI research perspective, we apply and ground our research methodology on theoretical frameworks from social sciences [43, 93, 97].

Central to our work is Activity theory [23, 48, 95], introduced in the following section, which defines how human activity is socially and culturally situated. We adopt Activity theory to define dialogues as purposeful activities and refer to them from now onward as *dialogue activities*. Activity theory also provides the definition of breakdown situations as applied in this work. Other central theories assist in defining how people share intentions [93] (Section 2.2), and how they communicate in dialogue activities [97] (Section 2.3).

2.1 Activity Theory

Activity, as a theory, was developed by Leontiev [47, 48] and has its foundations in socio-cultural psychology. This socio-cultural perspective on psychology rose from the efforts of two notable psychologists, Vygotsky and Rubinstein. They argued that the human mind is inseparable from society and culture. Rather, human activities within society and culture act as generative forces for creating the mind.

By employing the different ideas of Vygotsky and Rubinstein [25], Leontiev defined *activity* as an interaction between the *actors* (in our work, embodied agent

and human as *subjects*) and the world (*objectives* or *objects* of the activity). The actors and the world interact using *mediating artefacts* (tools) such as language. The *object* of activity is a central concept in Activity Theory, as it defines, motivates, and coordinates the activity. Furthermore, objects related to the subjects' motives and interests address their needs. In this work, we assume that an embodied agent comes with its own agenda to be fulfilled when collaborating with humans and, consequently, is referred to as a *subject.* Furthermore, assuming that the embodied agent is designed to serve the human, it adopts human motives to address the needs of the human.

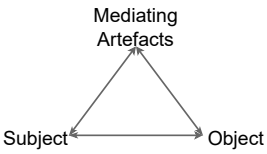
The *subject-object* interaction mediated by artefacts can be characterised as purposeful, transformative, and evolving [47, 48]. Leontiev's view of activity is illustrated in Figure 2.1a. As we can see, Leontiev's conceptualisation of activity was mainly in the context of an individual, and a conceptual framework for collective activities was missing. Engeström addressed this gap by introducing an extended Activity System model [22], also known as Engeström's Triangle. In the remainder of this thesis, we refer to this model as Engeström's Activity Model or EAM.

EAM provides a conceptual framework for activities performed by collectives. EAM is an extension of Leontiev's activity defined as *subject-object* interaction to an activity with an additional component, *community.* In the rest of the thesis, we refer to such activity a collective performs as *joint activity*.

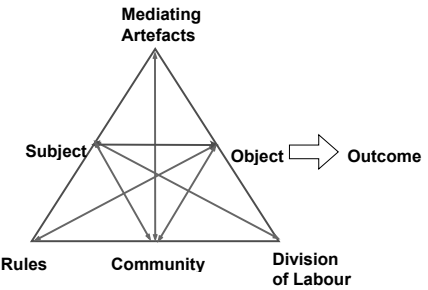
EAM also includes *rules* and *division of labour* to establish interaction between the community, the actors, and the objects, as shown in Figure 2.1b. In our work, humans and embodied agents constitute the community. The explicit and implicit rules facilitate the interaction between the subject/s and the community to perform the joint activity. The subject uses division of labour to coordinate their actions with other community members to realise the joint activity, relating the object with the community. The *outcome* in EAM integrates the actions' impact on the joint activity as contributing to social and cultural transformation.

The conflicts in EAM are represented by the *contradictions*, which may lead to breakdown situations. These conflicts lead the focus to shift from the central object of the activity to a different focus, for instance, correcting the artefacts or modifying the rules. These conflicts, however, are not seen as problems in Activity Theory but as opportunities to learn something new and develop [7].

Activity Theory has been applied as an analytical tool to study and develop an understanding of how socio-cultural aspects affect human psychology and development [79] in education, healthcare, organisational learning [23, 24], and in Human-Computer Interaction (HCI) [42, 43]. For an overview of how Activity Theory evolved over almost a century, see [79].



(a) Representing Leontiev's individual activity between the subject and the object facilitated by mediating artefacts.



(b) Engeström's Activity Triangle representing joint activity.

Figure 2.1: (a) Represents Leontiev's Activity of an individual and (b) Representation of a Joint Activity by Engeström.

Researchers have adopted Activity Theory also to define embodied agents [40, 65]. For example, in [65], the authors provide a conceptual framework for coordination between humans and embodied agents based on Activity Theory. They defined this coordination as a transition between distinctively defined subjective and objective coordination. During subjective coordination, actors cooperate to fulfil a goal based on their perception and understanding of other actors and the environment. In objective coordination, actors use prior and perceived knowledge to fulfil an observer's goal.

In a similar line of research but on human learning in collaboration with multi-agent systems (MAS), Liang and colleagues [49] apply EAM, focusing on pedagogical scenarios. Another recent application of Activity Theory in determining social and intelligent embodied agents for pedagogical scenarios is by Dolata and colleagues [19], where they evaluate a model of activity learning based on Activity Theory. The model assessed a pedagogical activity's characteristics and learning outcomes. The evaluation resulted in design recommendations for embodied agents to be used in pedagogy.

Closer to the focus of the scenario in this thesis is to apply Activity Theory to healthcare [33, 34]. In [34], authors combine concepts from Activity Theory and Argumentation theory reasoning [20] to provide tailored health-related activity recommendations to users. Activity Theory was used to develop an understanding of human activity and behaviour. In [33], authors use Activity Theory and co-design methodology involving older adults and caregivers to inform the design and development of a theoretical framework for personalised healthcare support.

Leontiev describes human activity as organised hierarchically at three levels. At the highest level is the *activity* aligning with a *motive*, which fulfils the needs of an individual. Levels of conscious processes called *actions* managed by goals, organise the activity. At the lowest level are internalised units of activity, called *operations* conducted unconsciously, without their own goals, restricted by *conditions.* Figure 2.2 illustrates different units of activity organised as a hierarchy.

The 'hierarchy of activity' in Activity Theory [48] was used in [33] to represent an agent's knowledge about human activities.

Similar to the work in this thesis, the works of [33] also defined breakdown situations for healthcare scenarios. Another relevant work that studied breakdown situations from an Activity theoretical perspective is [75]. They perform a thematic analysis and present causes for breakdown situations in interactions between an embodied agent tutoring elementary school children. Their results listed breakdown situations due to controller errors, inconsistent behaviour, lack of fairness, inability to understand communication issues, and evoke engagement.

Building upon previous works, EAMs, the hierarchy of activity, and the concepts of focus shifts and contradictions, this thesis studies dialogue activities and embedded breakdown situations in day-to-day healthcare dialogues.

We interpret a dialogue activity as an EAM. The contradictions are seen as causes of breakdown situations in the dialogue activities. This thesis focuses on breakdown situations when the mediating artefacts (the tools) fail to work or are unsuitable; there is insufficient or inconsistent knowledge about rules, division of labour, or the motive; and when there are conflicting objects.

The object is interpreted as *intention* in this thesis, highlighting the purpose of the subject and other actors in the community. To define the internal relationship between the intention, the subject (human actor and embodied agent), and the community, we explored the theory of *We-intention* [93]. We-intention defines intentions to perform joint activities. We use intention as the object of activity in general and We-intention to denote the *object* being co-created in a dialogue-based joint activity.

2.2 Collective Intentionality

Humans are considered intentional actors driven by their needs. From a philosophical point of view, intentions can be designated to individuals or groups. Collective intentionality is an umbrella term to study intentions ascribed to groups.

Philosophers and social scientists have expressed collective intentionality as *shared intention, collective acceptance, joint attention, We-intention, joint commitment, and collective emotion.* Shared intention enables coordinated and cooperative

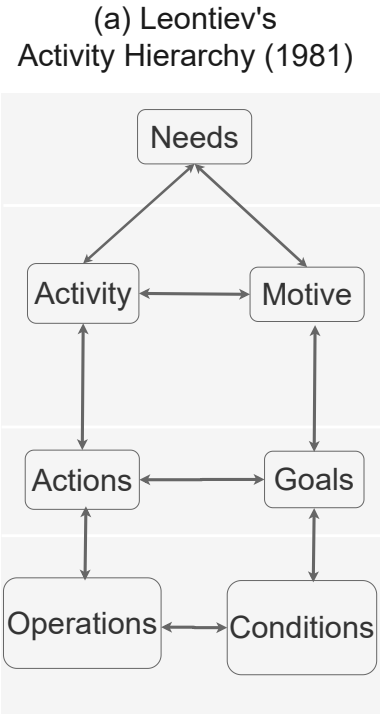


Figure 2.2: Hierarchy of Activity composed of motive-driven activity at the highest level, goal-directed actions, and conditional operations at the lower levels.

actions to achieve collective goals [9]. Collective acceptance allows the existence of institutions, symbols, and language [92]. Joint attention establishes common ground to perform cooperative activities [90]. Collective emotion provides us with what a group accepts and prepares them for collective actions [94].

Specific to our research work is the conceptualisation of collective intentionality as We-intention by Sellars [74]. Sellars proposes We-intention as *an attitude with a shared point of view* [74] assessing each others' contributions and involves identification to a group. Sellars categorises We-intention into two types– primary I-referential, which contributes to individual actions, and we-derivative I-referential, which contributes to joint actions. Sellars argues that many intentions are we-derivative; for instance, if a football goalkeeper prevents a goal from happening, it is a derivative of our shared intention to play the game of football as a team. He

adds that to intend x, actors must do their part or share of actions. If an actor shares fewer beliefs and intentions, seeing them as *"one of us"* becomes difficult (*[72]. Pp. 203*).

Tuomela and Miller's conceptualisation of Sellars's notion of We-intention form the theoretical basis for defining the object of dialogue activity, and our computational frameworks attempt to model them.

2.2.1 We-intention

We-intention are *attitudes* that actors have while performing actions as a collective. These attitudes differ from the case that they were forced or acted out of fear. We-intentions do not get things done, and in the works of Tuomela [91], they are *aim-intentions* in which the actor is *"assumed by his action to contribute towards the aimed result" ([91]. Pp. 37)* and the actors see to it that something be the case.

A We-intention can be expressed with the following schema [93]: (W1) Where involved actors who recognise themselves as part of a community C commit to realising X by performing their role.

We-intention contrasts with joint intentions, an *action-intention* [91]. In joint intentions, the actor performs actions and believes that with some non-zero probability can achieve the goal by their action/s. Such actions are called *joint actions,* which can be performed jointly with others but can also be done individually [93].

Actors can have a We-intention and reason about a joint intention as a group or private individual. When actors act as a group on an agreed-upon common goal, they are intending and working in a *We-mode* contrary to *I-mode* where actors act individually [91] motivated by their private goals.

Tuomela contrasts We-intention I-mode to be similar to Bratman's *shared intention* [91]. Bratman represents shared intentions as held by a group, referred to as *we*, where shared intention only holds if all the members in *we* have the same intention. This intention can be achieved by interlocking all the plans and having common knowledge among the members [9].

To co-create We-intention, the actors and the community require mediating artefacts that can be language or gestures. Our work focuses on language-based artefacts. The underlying theory we explored to define the mediating artefact is argumentation-based dialogue theory [97], which provides different types that can be used to conduct dialogues.

2.3 Argumentation-based Dialogues

Using dialogue as a context to evaluate arguments was first proposed by Hamblin in his research work *Fallacies* [36]. An under-developed yet central concept in Fallacies was *commitment*, which was later developed and refined to evaluate arguments using dialogues by Walton and Krabbe in their seminal research work *Commitment in Dialogues* [97].

Walton and Krabbe develop their theory for argumentation-based dialogues around one type of commitment that is associated with performing an action, where collectives (such as organisations, states, unions) or actors, A are committed to a course of action, X, forming a commitment bond, represented by the following schema: A is bound to X. These commitments are mediated in dialogues with other actors. Depending on the commitment, Walton and Krabbe define seven types of dialogues that actors can apply to mediate their commitments [96].

We focus on four different dialogue types: *persuasion, deliberation, inquiry,* and *information-seeking.* These dialogue types are characterised by an initial situation, the main goal, and internal goals.

- **Persuasion-** emerges from an initial situation of a conflict due to a difference in the views of the involved actors. The main goal is to resolve the conflict. The actor's internal goal is to convince and persuade the other to take their point of view. Persuasion is a sequence of attacks, defences, or questions and replies called *moves*, which are locutionary sequences made by actors taking turns in the dialogue activities.

- **Inquiry-** emerges from a general ignorance or a need to build new knowledge. The main goal is to widen and increase the knowledge and achieve a stable agreement between the goals and proposition that answers the question. For the inquiry to succeed, actors must subscribe to one conclusion. If an actor concludes earlier than the others, they must convince the others by providing the basis for the conclusion.

- **Deliberation-** is similar to inquiry in that it starts with an open problem, to agree on a plan of action to be influential to the outcome in some way.

- **Information-seeking-** emerges from one actor possessing some information that the other one lacks and needs, a kind of ignorance. The main goal is to spread knowledge, and the actors' internal aim is to gain, pass on, show, or hide knowledge.

To summarise, we have described the three foundational theories applied in this research. In the activity-theoretical sense, communication between human actors and embodied agents is defined as *dialogue activities*. *Breakdown situations* are described using the concept of contradictions. What the actors aim to achieve during a dialogue-based joint activity is the object interpreted as *intention*. Specifically, We-intention is used as a theoretical basis for defining intentions for dialogue activities. Finally, the actors conduct dialogue activities using mediating artefacts to communicate, interpreted as argumentation-based dialogue types, besides language, knowledge, and body.

Chapter 3

Studies of People's Perception and Expectations of Breakdown Situations in Dialogues

3.1 Problem Specification

In a dialogue-based joint activity, actors (human and embodied agent) generally cooperate with each other to co-create mutual understanding about an intention, and not on mediating artefacts such as language or what others ought to do (division of labour). However, different problems in co-creating understanding and conflicts may move the focus from creating We-intention to fixing the artefact or clarifying the procedures causing breakdown situations. This part of the work focuses on exploring people's *perceptions* and *expectations* of human-agent dialogue activities, including breakdown situations. We conducted user studies to understand the perception and expectations relating **RQ1** and **RQ2**, respectively. The setup is common to the three user studies in papers I, II, and V.

3.2 Our Approach

An activity analysis of fictive but realistic healthcare scenarios involving human-agent dialogues using Activity Theory was conducted. The scenarios were partially based on [83]. The outcome is illustrated in Figure 3.1. Based on the analysis, we designed dialogue activity scenarios embedding breakdown situations caused by various reasons, further described in Subsection 3.2.1. A user study to evaluate these scenarios was performed. The study setup is described in Subsection 3.2.2.

3.2.1 Construction of Dialogue Scenarios

In this thesis, we identified the following reasons for breakdown situations: when the mediating artefact is not working or is unsuitable; the knowledge about procedures, rules, and intentions must be clarified; and when conflicts between intentions arise. How these breakdown situations can be represented using the concepts in Activity Theory is illustrated in Figure 3.2. Breakdown situations move the focus from the central EAM to alternative EAMs. The following summarises the reasons for breakdown situations central to this thesis.

- **Insufficient or inconsistent factual knowledge-** when there is a lack of or contradictory knowledge about the facts, such as names of other actors, community members, or objects. (Indicated in Figure 3.2 by A)
- **Insufficient knowledge about intention-** when there is a lack of knowledge about why other actors are pursuing a specific intention. (Indicated in Figure 3.2 by B)
- **Insufficient or inconsistent procedural knowledge-** when there is a lack of or contradictory knowledge about the roles and tasks of others. (Indicated in Figure 3.2 by C)
- **Conflict of intention-** when there is a conflict between actors about which intention to pursue. (Indicated in Figure 3.2 by D)
- **Insufficient knowledge of social norms-** when there is a lack of knowledge of necessary social norms to interact with other actors. (Indicated in Figure 3.2 by E)

We designed six dialogue activity scenarios embedding breakdown situations for managing health and well-being. The scenarios are enacted in a home environment where an embodied agent assists human actors with care needs due to health

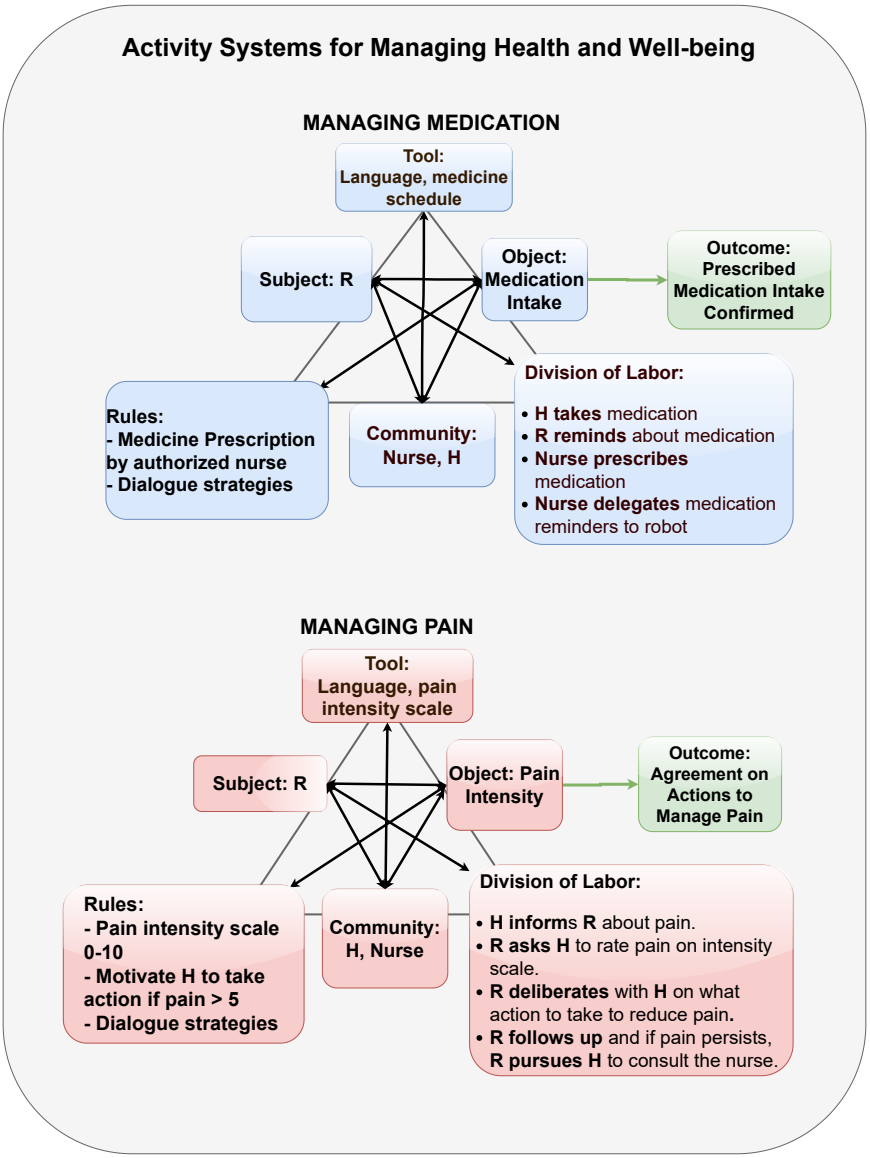


Figure 3.1: EAM to manage pain and medication to fulfil human actor's need to maintain health and well-being. *(Image Source– [86], Figure 1. Pp. 4.)*

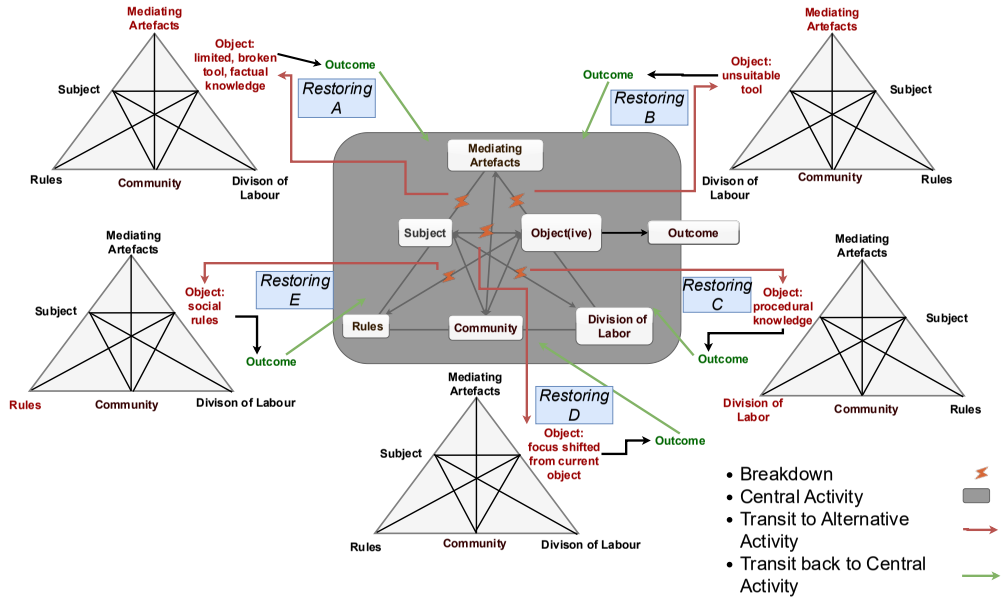


Figure 3.2: Illustration of breakdown situations (orange symbol) resulting in a shift in focus (red arrow) from the central EAM (enclosed in the Grey box) to a supporting EAM and then back after the situation is resolved (indicated by a green arrow). (*Image Source– [86]. Figure 2. Pp. 4.*)

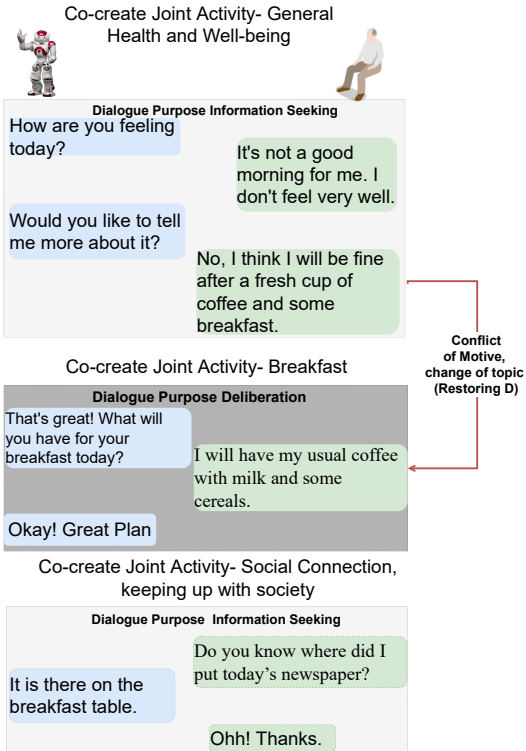


Figure 3.3: A fictive yet realistic dialogue between a socially intelligent agent and a human as part of a morning routine. (*Image Source– [86] Figure 3. Pp. 9.*)

conditions. One of the dialogue activity scenarios is illustrated in Figure 3.3, where a conflict of intention between the socially intelligent agent and the human is illustrated using a fictive dialogue designed using theory-based analysis. The agent starts the dialogue by asking how the human is feeling. To which the human indicates they are not feeling well. The proposed topic is rejected when the agent asks if they would like to talk about their well-being— resulting in a *situation of conflict.* The situation was resolved by the agent adapting to the topic introduced by the human about having breakfast. For details about the rest of the five scenarios, refer to Paper II [86].

The scenarios further embedded dialogue strategies based on argumentation-based dialogue types to manage the identified reasons for breakdown situations, as illustrated in the following:

1. Align intention by applying *deliberation* dialogue type with the human actor to manage conflict, given the embodied agent's intention is judged to be less important.

2. Apply *inquiry* and *information-seeking* dialogue types for creating new knowledge and addressing the lack of knowledge.

3. Apply *persuasion* dialogue type when there is a conflict of intention, and the embodied agent's intention has to be prioritised.

   Other strategies not part of argumentation-based dialogue types were:

4. *Apologise* before interrupting if the human actor is occupied with another activity.

5. *Repeat* the utterance when there is noise, and the human actor is not responding to the initiated dialogue activity.

3.2.2 Setting up User Studies

To prepare the user studies, the dialogue activity scenarios embedding breakdown situations and strategies defined in Section 3.2.1 were enacted in a Wizard-of-Oz (WoZ) setup [31] between two volunteers and an embodied agent.

Wizard of Oz Recordings

One older female and a younger male volunteered to interact with the embodied agent in a WoZ setup. Both volunteers provided consent to be recorded. Volunteers interacted separately with an embodied agent (a Nao robot[1]) in six sessions each. The interactions were audio and video recorded.

Participants

Participants were recruited through convenience and snowball sampling on social media networks. The study participants were required to have intermediate to fluent knowledge of the English language and be able to provide consent and participate in video conferencing using Zoom. A total of twenty participants participated in the

[1] https://www.aldebaran.com/en/nao

studies. Of these participants, there were twelve women and eight men in the age range of 23-72. Eight participated in Study I (Paper I), 20 in Study II, including the first eight (Paper II), and Study III (Paper V) included a focused analysis of three dialogue activity scenarios and the data collected from 20 participants.

Procedure followed in the User Studies

Data collection was conducted remotely due to the pandemic. Participants watched video recordings and participated in a semi-structured interview over Zoom. The interview questions addressed participants' perceptions of the behaviour of the actors, their intentions, and how they understood the situation in the recordings. Furthermore, we asked participants about the perceived and expected roles of the embodied agent in the interaction, the limitations of interactions involving embodied agents, and their obligations in a healthcare scenario. These interview questions are detailed in [86]. The interviews were audio and video recorded and transcribed verbatim.

Data Storage

The collected data was stored locally on the personal computer of the primary researcher. The exchange of data for transcription and analysis with other involved researchers was done through secure university channels. The transcribed data was anonymised, gathered and stored locally on the personal computer. After finishing the analysis, a single copy of the transcribed anonymous data was retained with the primary researcher.

Data Analysis

The transcribed interview data set was analysed using thematic analysis [14]. Thematic analysis is a method in which *themes* are derived to provide a rich structure and description to the data set. Thematic analysis was performed by (1) familiarising with the data set as closely as required; (2) highlighting and coding the text in the data set that fulfilled the research work's objective; (3) comparing and discussing main codes and deriving themes; and (4) collating and finalising the themes.

The thematic analysis was performed, providing results to papers I [87], II [86], and V [35]. The respective results are detailed in each of the papers and summarised in Section 6.1, Section 6.2, and Section 6.5.

Chapter 4

Recognising Breakdown Situations in Dialogues Relating to Syntax, Semantics, and Social Aspects

4.1 Problem Specification

In Chapter 3, we identified and defined various reasons for breakdown situations. Including those due to faulty mediating artefacts, the conflict between intentions, to lack of or contradictory knowledge about the facts, procedures, intentions, and norms. These reasons for breakdown situations demand creating understanding of the different aspects, such as the syntax and the semantics used in dialogue activities. Creating methods enabling embodied agents to develop such an understanding is an open and complex research problem [16]. This part of the research explores methods for recognising breakdown situations relating to **RQ3,** corresponding to papers III and IV.

4.2 Our Approach

We interpret the creation of an understanding as meaning formation about the formal properties, categories of events and other actors, intentions and interests of the self

and other actors, how the intentions are related to each other, what are the cultural context, and the ultimate motive of an actor behaving in a certain way [80]. When interpreted for dialogue activities, the meaning formation corresponds to the syntax, semantics, objective, social, and normative aspects as described by Steels [80].

We consult and embed theories about syntactic, semantics, and social aspects of dialogue activities (described in subsections 4.2.1 and 4.2.2) into our computational frameworks. We make assumptions, build our computational frameworks, and apply methods to evaluate them.

4.2.1 Computational Framework to Recognise Breakdown Situations from Syntactic Relationships

To create a computational framework for understanding the syntax of breakdown situations, we explored the concepts of *"quantity"* and *"structure"* relating to words.

The quantity of words in a dialogue activity relates to the *amount of information* provided. Grice has conceptualised this idea about the amount of information in his work Cooperative Principle (CP) as *Maxim of Quantity (MoQ)* [32].

The structure of words implies how words (often referred to as lexical items) are related to each other in a dialogue activity turn. We explore this relationship of words to build a symbolic computational framework representing MoQ.

Subsequently, we briefly describe Grice's CP and MoQ, present our assumption, and describe how we developed and evaluated the framework.

Theories and Methods

Grice, in his work [32], proposed utterances in dialogue activities to have dual meaning, one that can be derived from what the words 'mean', that is their semantics and another that can be 'implied' from what was said. Grice has defined this implied meaning as falling into either the conventional meaning, general semantics or non-conventional meaning, framed within the context of an ongoing dialogue activity.

Grice provided a general principle for optimal dialogue activities: *"make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged."([32]. Pp. 45)*. Grice states that if we accept such a general principle, we can distinguish four categories under which maxims (rule of conduct) can be defined to yield Cooperative Principals (CP) with the following categories – *quantity, quality, relation,* and *manner.*

To develop the understanding of syntax in breakdown situations, the category of *quantity MoQ* was explored, which relates to the amount of information being provided and has the following maxims embedded in MoQ:

- Provide as much information as is required and be optimally informative.
- Do not provide more information than is required.

To develop a computational framework representing MoQ as *syntactic understanding*, we interpret MoQ as *informativeness* of a turn in a dialogue activity. Informativeness is a score derived from a single turn's *structure between words.*

We use a concept of *syntactic cohesion* to derive the informativeness score. We follow the definition of cohesion in [17] to define syntactic cohesion as *"that establishes relationships between presupposed and presupposing words, such as conjunctions, interjections, substitution, reference, and ellipsis in a text" ([84]. Pp. 1).*
Therefore, we explore our assumption that *informativeness is intrinsically related to syntactic cohesion.*

The method for developing the framework includes dependency parsing. Dependency parsing is a technique that finds relationships between the words, such as subject, object, and modifier. These relationships are called *dependencies*. Dependencies usually follow a nesting structure related to the linear order, hence forming projective dependency trees [60]. In these dependency trees, the dependencies are established with arrows between words. The word from which an arrow originates is called the *head*, and the words on which the arrows point are *dependants*. A head can have one or more dependants.

A computational framework embedding the syntactic cohesion and informativeness scores was developed.

Evaluating the Assumption

For evaluating our assumption, informativeness and syntactic cohesion scores were computed on a dataset called TeamTalk Corpus [53]. This dataset consisted of navigation instructions (1600) that humans fluent in English provided to embodied agents. The relation between the two developed scoring methods and other features was analysed. The scores provided by the computational framework were compared with human scores provided during a survey on a subset of the dataset. The code is publicly available on GitHub [1]. The results are presented in Paper III [84] and summarised in Section 6.3.

[1] https://github.com/maitreyeeT/Informativeness

4.2.2 Developing an Annotation Schema with Semantics and Social Aspects

Previous research work [85] observed that breakdown situations often span several turns, requiring methods to capture such long sequences.

We approach the problem of capturing long sequences by adopting different theories relating to semantics and the social aspect of dialogue activities. The theories are embedded to develop an annotation schema. We form an assumption on our annotation schema and develop computational methods to evaluate the scheme.

Theories and Methods

We explore *dialogue acts* [12] to represent the semantics and *sequence expansion* [81] for the social aspect of dialogue activities.

Dialogue Interpretation Theory: According to [11], a dialogue act has two main components: semantic content that provides information about the type of dialogue act and communicative function that specifies how the actors update their understanding of the information and update the information state(s) of others involved. For developing the annotation schema, communicative functions were used. There are seven categories of communicative functions constituting 56 types (for a detailed description of all the functions, refer to [12]).

Conversation Analysis: The socio-linguistic domain of Conversation Analysis (CA) studies the sequential organisation of everyday dialogue activities or any social interaction. This sequential organisation is assumed to represent actors' understanding of each other, forming a social interaction [41].

At the level of turns, CA [76] defines dialogue activities as a composition of sequentially produced turn construction units (TCU), which can be of a verbal or non-verbal nature. TCUs share a reflexive relationship with their prior and later counterparts. TCU in its base form is referred to as adjacency pairs [68] consisting of *first-pair* part and a *second pair* part. Some common adjacency pairs are– greeting-greeting, question-answer, offer-accept, and offer-reject.

For our purposes, we want to go beyond base form (adjacency pairs) and be able to capture longer sequences of turns. CA provides the concept of sequence expansion to develop such capturing mechanisms [81].

Sequence expansion: Within the field of CA, sequence expansion describes sequences that can expand a base form. There are three forms of such expansions: *pre-expansion, insert expansion,* and *post expansion*.

In a dialogue activity, an actor applies a pre-expansion to prepare another actor for an upcoming base form. An insert expansion can be applied to resolve issues concerning the first-pair part or to clarify circumstances determining the second-pair

part. Post expansions are applied to resolve issues related to the second-pair part, ask for additional information, or initiate a closure of the dialogue activity.

Our assumption relates to how combining different aspects of a dialogue activity would potentially fulfil our purpose of capturing longer sequences.

Therefore, *the annotation schema embedding information about syntax, semantics, and social aspect could be used for capturing longer sequences of turns.*

Evaluating the Assumption

The annotation schema was applied to label a dataset. The dataset consisted of 89 dialogue activities gathered from publicly available sources [12, 37] and others synthetically created or gathered by transcribing natural dialogues. The labelled dataset provided input to learning-based computational methods. Comparison and evaluation of baseline and main methods were performed. The models and the annotated dataset are publicly available on Github [2]. The results are presented in Paper IV [89] and summarised in Section 6.4.

[2] https://github.com/maitreyeeT/Seq2SeqDialStruct

Chapter 5

Recognising and Managing Conflict of Intention in Dialogues

5.1 Problem Specification

This part of the thesis explores how agents can recognise and manage breakdown situations. The focus is on the breakdown situations caused by *conflict of intentions.* A conflict can occur when the actors prioritise different intentions rather than co-creating one. This was explored in papers V and VI, relating to **RQ3** and **RQ4.**

5.2 Our Approach

We approach the problem of recognising and managing conflict of intentions by adopting theories from psychology and the social sciences, illustrated in Subsection 5.2.1. Reasoning techniques were explored to recognise and manage breakdown situations, further explained in Subsection 5.2.2 and Subsection 5.2.3.

5.2.1 Theories and Methods

In Theory-Theory [73], researchers argue that human actors develop a theory to predict and explain other actors' behaviour caused by mental states. In Simulation Theory [30], human actors generate similar states and processes in themselves,

representing a simulation of psychological states and processes of other actors to understand those processes and states associated with the simulated target actor. These theories are commonly referred to in the research community as mentalising, mindreading, and Theory of Mind (ToM). We apply ToM to understand how human actors ascribe mental states to other actors and reason about the behaviour of the others. ToM helps us develop mechanisms for agents allowing reasoning and decision-making that considers, for instance, human beliefs, desires, needs, and expectations.

A presupposition in ToM is that the actors have mental states. One approach to defining agents with mental states is the *Belief-Desire-Intention* (BDI) framework [8, 64]. We apply BDI to characterise our embodied agent with mental states of beliefs, desires and intentions. BDI frameworks were initially defined for individual agents. However, researchers have also applied the BDI framework to build agents that perform joint activities [3, 15, 18, 44, 98]. In this thesis, we follow a similar line of research and develop BDI agents capable of participating in joint activities.

5.2.2 Decision-making using Non-monotonic Reasoning to Manage Conflict of Intentions

We define a computational framework in Paper V to realise ToM-like reasoning about the breakdown situation due to conflict of intentions. The framework involves extending the BDI framework representing the actors and applying *Answer Set Programming (ASP)* for reasoning.

ASP is a rule-based language belonging to the logic programming paradigm, which allows *non-monotonic* reasoning by generating programs. In non-monotonic reasoning, the programs generate tentative conclusions. They are subject to change depending on the availability of new evidence, which is a main advantage compared to traditional logic frameworks that assume all the information exists.

To cater for the dynamic nature of dialogue activity embedding conflict of intentions, ASP is a strong candidate for reasoning. To perform reasoning, ASP has a syntax represented by symbols and rules and semantics to interpret those rules. For more details on the syntax and semantics of ASP, we direct the readers to [5, 59].

5.2.3 Decision-making using Probabilistic Reasoning to Manage Conflict of Intentions

We present another computational framework to recognise and manage conflict of intentions in Paper VI by planning and executing plans aligning always with

human intentions. This research focuses on developing a framework for agents to manage a conflict of intention by adapting to humans intention. The research presented in Paper VI builds upon our earlier research in [62]. Where we provided an initial formulation of We-intention and a dialogue state-machine for mediating the intention.

In this part of the research, the initial formulation of We-intention in [62] is extended with a ToM-like reasoning to distinguish between We-mode and I-mode We-intentions. The framework involves two BDI models representing the agent itself and the other actor. The agent performs probability-based reasoning and creates plans using hierarchical task networks (HTN).

HTN is a planning method to act in an environment where the objective is not to achieve a goal but to perform tasks under constraints. The constraints impose an order in which the tasks need to be completed. These task networks can either consist of primitive tasks that can be executed directly or compound tasks that have partially ordered sub-tasks [27].

Chapter 6

Summary of Contributions

In this Chapter, we describe the aim of the papers included in this thesis. Then, we describe how their research contributes to addressing this thesis's research questions. We conclude by presenting an initial formulation of a cognitive architecture embedding the contributions of this thesis.

6.1 Paper I

Maitreyee Tewari and Helena Lindgren. Younger and Older Adults' Perceptions on Role, Behavior, Goal, and Recovery Strategies for Managing Breakdown Situations in Human-Robot Dialogues. *In Proceedings of the 9th International Conference on Human-Agent Interaction, ACM Digital Library, Pp. 433-437, 2021.*

The research presented in Paper I explored how young and old participants perceive actors' roles, goals, behaviours, and strategies in dialogue-based joint activities embedding breakdown situations. The aim was to provide guidelines for developing methods that enable embodied agents to manage dialogue activities, including breakdown situations.

The user study presented in this paper follows the study set-up described in Chapter 3. A subset of the gathered data (interviews of eight out of 20 participants) was thematically analysed.

This paper's first major contributions are the following three themes derived from the analysis of the user study data: *roles and relationships, understanding and emotional connection,* and *adaptive behaviour and manner.*

The following is a summary of the results from three themes:

1. **Roles and relationships-** participants found the embodied agent played different roles, from being equal, such as a companion or assistant, and other times more authoritative, like a nurse. Participants expected the embodied agent to act as a *companion* and provide emotional support and somebody who listens to them; be proactive in pushing the person into following healthy routines and monitor health, acting like a *care assistant.* People expected that humans, when interacting with such agents, must have patience and assist them in learning to build and establish long-lasting relationships.

2. **Understanding and emotional connection-** the embodied agents' understanding of breakdown situations was found mostly intuitive apart from when it lacked the understanding of volunteers' emotions and unwillingness to interact. The embodied agent was expected to express compassion and proactively ask for the appropriateness of its behaviour.

3. **Adaptive behaviour and manner-** the embodied agent was perceived as polite, caring, cooperative, and pleasant. An exception was when it lacked the ability to adapt to the needs of an individual and the situation. It was highlighted that adapting to a person's needs and situations is essential for an embodied agent.

The second major contribution is the design implications based on the themes. The following is a summary of the design implications:

1. **Co-constructing knowledge and learning-** the embodied agent needs to have a model embedding the knowledge about itself (for example its roles and goals), other actors, and the activity to be performed, including the motives, domain knowledge, rules, norms, and the division of labour.

2. **Co-constructing understanding and creating dynamic memories-** the embodied agent needs to understand the reasons for and have strategies to manage breakdown situations. Paper I focused on identifying and managing breakdown situations requiring knowledge adjustment and co-creating intentions. For knowledge adjustment, the strategies consisted of: (a) in case of misunderstanding, stating inconsistent knowledge and asking for confirmation; (b) in case of non-understanding, stating the insufficient knowledge and asking for more information; (c) when a non-hearing occurs, then repeating the information slowly and loudly. When co-creation of intentions cannot be accomplished due to noncooperation or focus shift, then (A) collaborate on

the topic introduced by the human actor, (B) if the agent's topic is important then persuade the human actor, and (C) manage two parallel topics.

3. **Natural speech generation-** the embodied agent should avoid long sentences and multiple topics within a turn and provide as much information as required.
4. **Planning and decision making-** The embodied agent should facilitate human participation in decision-making and apply interpretable and transparent computational processes.
5. **Self-reflecting-** The embodied agent should rationalise the effects of its potential actions on human autonomy, agency, and emotional states.

The themes provided an increased understanding of dialogue activities. Thus, addressing the research questions **RQ1** and **RQ2**. The design implications guide the development of human-centred mechanisms for human-agent dialogue-based joint activities addressing research questions **RQ3** and **RQ4.** Interesting to explore in the future is how these implications can be embedded and implemented in computational frameworks.

6.2 Paper II

Maitreyee Tewari and Helena Lindgren. Expecting, Understanding, Relating and Interacting - Older, Middle-aged, and Younger Adults' Perspectives on Breakdown Situations in Human-Robot Dialogues. *Frontiers in Robotics and AI, Frontiers Media SA, Volume 9: 956709, 2022.*

The aim of the research presented in Paper II was to study how people perceive breakdown situations caused by insufficient and inconsistent knowledge of different types, such as norms and procedures and conflict of intentions. Furthermore, how people experience an embodied agent's strategies to manage breakdown situations was explored.

The user study presented in this paper followed the set-up described in Chapter 3. Data collected in interviews with 20 participants were analysed using thematic analysis. To increase the credibility of our findings, a triangulation approach was undertaken, where the primary researcher and a Master's student performed their analysis individually and discussed them.

The analysis of user study data resulted in themes and taxonomy about aspects of human-agent dialogue-based joint activities. In the following, we briefly describe the contributions that address the research questions of this thesis and an additional contribution of methodology development.

1. Themes increasing our understanding of people's perceptions about the *reasons* for breakdown situations and *strategies* to manage those.

2. The analysis resulted in an over-arching theme of *expecting*. The expecting theme consisted of three sub-themes of *understanding, relating,* and *interacting*. The aspects the agent needs to create an understanding for detecting breakdown situations were provided. The *interacting* theme suggested ways to use gestures and body movements naturally to prevent breakdown situations. The theme of *relating* provides recommendations on how the agent should relate to the human and interchangeably play the roles of a companion, an assistant, and a professional.

3. A hierarchical taxonomy for breakdown situations organised as three categories of factors:

   (a) Factors affecting the *understanding* related to what caused breakdown situations. The following reasons were highlighted: insufficient knowledge about how the human was feeling, their mood, the ongoing activity, procedures such as who does what, and social norms, while other reasons related to inconsistent behaviour and conflict of intentions.

   (b) Factors affecting the *interaction* related to how the embodied agent interacted with the human using body and speech cues, behaviour and language. Participants found the embodied agent's behaviour was sometimes not empathetic and intrusive. They reflected on how the embodied agent looked around while talking to the volunteers and how the speech was slow, unclear, childlike, and lacked emotions, sometimes causing breakdowns.

   (c) Factors affecting the *relating* was how the embodied agent related with humans during the dialogue activity. Participants commented that the embodied agent acted as a subordinate, like an assistant, being equal, like a companion, and at other times was more authoritative, like a nurse or a teacher.

4. To meet the expectations for developing embodied agents highlighted the need to create an *understanding* about a situation in relation to the other actors' emotions, intentions, the norms to be followed, necessary facts, and the roles of the self and other actors. When interacting, the embodied agent must organise the body with appropriate speech, behaviour, and response. In the case of older human actors, the embodied agent should be able to interact in their native language. The *relating* implication emphasises embodied

agents be socially adaptive, establish, co-create, and maintain relationships and interchangeably enact different roles.

5. Activity Theory-based methodology to design human-agent dialogue activities embedding breakdown situations.

The research in this paper contributes to addressing the research questions **RQ1** and **RQ2.** The results contribute to developing a deeper understanding of dialogue activities and embedded breakdown situations. The design implications and research agendas guide the future exploration and development of human-centred computational mechanisms, that could allow to create deeper and improved understanding in agents about dialogue activities and breakdown situations.

6.3 Paper III

Maitreyee Tewari, Suna Bensch, Thomas Hellström and Kai-Florian Richter. Modelling Grice's Maxim of Quantity as Informativeness for Short Text. *In ICLLL 2020: The Proceedings of the 10th International Conference in Languages, Literature, and Linguistics, Pp. 1-7, 2020.*

The aim of the research presented in Paper III was to create a computational framework for recognising breakdown situations using the syntactic aspect of dialogue activities. This addresses the design implication *3. Natural speech generation* in Paper I to provide as much information as is required during dialogue activities.

A novel computational framework was developed that represents Grice's CP *Maxim of Quantity (MoQ)* as described in Chapter 4. This was realised by extracting syntactic dependency relationships between words and scoring their *informativeness,* which tells the amount of information provided in a dialogue activity and *syntactic cohesion,* describing the relationship between words.

Each instruction dialogue in the TeamTalk dataset was scored with *informativeness, syntactic cohesion,* and *length* scores. The assumption about the relationship between informativeness and syntactic cohesion was evaluated using statistics and by conducting a survey with 19 participants. Scores provided by the participants were compared with those generated by our computational framework.

The results showed high correspondence between participants' scores and those provided by our framework. The feature analysis and descriptive statistics confirmed our assumption about the intrinsic relationship between informativeness and syntactic cohesion.

This paper addressed **RQ3** by developing and evaluating recognition mechanisms related to lack or excess information. In the future, the framework can be

adapted for an embodied agent to respond with optimal information, preventing breakdown situations. Furthermore, it could be used to categorise the actors' responses and predict breakdown situations related to inconsistent or insufficient knowledge. Agents could employ such methods to reflect on past interactions to avoid breakdown situations from occurring in the future.

6.4 Paper IV

Maitreyee Tewari and Michele Persiani. Variational Autoencoding Dialogue Sub-Structures Using a Novel Hierarchical Annotation Schema. *In Proceedings of the 6th IEEE Congress on Information Science and Technology (CiSt), IEEE, Pp. 334-341, 2020.*

The research in Paper IV aimed to create a computational framework that can understand the syntax, semantic and social aspects of breakdown situations spanning across several turns. The research builds upon our earlier works [82, 83, 85]. This paper contributes towards addressing the design implication *2. Co-constructing understanding and creating dynamic memories* in Paper I, and *understanding* in Paper II to recognise reasons for breakdown situations.

The contributions of Paper IV are the following:

1. An annotation schema representing a dialogue turn's syntax, semantics, and social aspects. We split a turn and call the resulting splits as *segments*. A segment here differs from the one we used in Section 4.2.1. For the annotation schema, we define a segment as the longest stretch of words or non-verbal elements that expresses a communicative function [13].

    (a) The organisation of turns is interpreted to provide an understanding of the social aspect of a dialogue activity. Therefore, the annotation schema captures information about a turn's social aspect at the highest level, using labels representing sequence expansions and base form explained in Chapter 4. At the same level, the annotation schema also represents the social aspect using the dialogue policies, defined in [82]. The three dialogue policies were: (1) *binding policy,* which restricts the actors to apply a specific communicative function, for example, a question or a request binds another actor to respond by answering, confirming or rejecting respectively; (2) *progressive policy* invites the actors to change the topic or conclude the dialogue activity; (3) *co-occurring policy* enables the actor to use communicative functions together with a base form part. In co-occurring functions, the other actor doesn't need to

respond to each function; for example, an answer can be accompanied by stalling and feedback functions.

(b) At a lower level lies the aspect of the semantics of a turn. In the annotation schema, the *semantics* is represented by communicative function labels as described in Chapter 4.

(c) At the lowest level are the syntactic aspects of a dialogue activity. The annotation schema represents the *syntactic* aspect using the same concept, *"structure"* of words as described in Chapter 4 and the technique of dependency parsing. The syntax is comprised of the following structural information about words in a turn: *subject-object-verb, auxiliary verb, noun, interjections,* and *adverbs.*

2. An annotated dataset was developed by combining manual and automatic annotation using the annotation schema. The proposed annotation schema was applied to label a dataset of 89 dialogues in two phases. In the first phase, Tewari manually labelled the dataset with communicative functions representing the semantics. In the second phase, the syntax was extracted using dependency parsing. The social aspect consisting of dialogue policies and sequence expansions was labelled using a rule-based method.

3. A computational framework consisting of partially annotating the dataset and a learning-based method to learn and generate the representation of different aspects. The labelled dataset enabled the learning-based computational method to generate long sequences of three, four, and five turns. Therefore, the annotation schema facilitates an agent to use knowledge from multiple turns instead of just one preceding turn.

This paper contributes to the research question **RQ3** by developing and evaluating a novel annotation schema that could be applied to recognise breakdown situations. An interesting future work will be on developing mechanisms for embodied agents to use the knowledge about sequence expansion to manage breakdown situations and dialogue activity.

6.5 Paper V

Esteban Guerrero, **Maitreyee Tewari,** Helena Lindgren and Panu Kalmi. Forming *We-intentions* under Breakdown Situations in Human-Robot Interactions. *Computer Methods and Programs in Biomedicine, Elsevier, Volume 242: 107817, 2023.*

Paper V aimed to develop computational mechanisms to reason when a conflict of intentions happens. This paper builds on and contributes toward addressing the following design implications in Paper I: *(1) co-constructing knowledge about other actors and the activity, (2) co-constructing understanding and creating dynamic memories* about reasons for conflicts and applying strategies to manage those, and *(4) transparent decision making.*

Paper V makes the following major contributions:

1. The definition of three strategies to *align, reject,* or *partially align* with other actors' intentions in situations of conflicting intentions. Aligning, rejecting, or partially aligning mirrors strategies A *(i.e. initiate cooperation)*, B *(i.e. negotiate by persuading)*, and C *(i.e. to continue with two parallel topics)*, respectively, in Paper I.

2. A novel computational framework representing the agent's and other actors' mental states and a reasoning mechanism to manage conflict of intentions.

3. An analysis of people's perception of conflict of intentions among actors (humans and an embodied agent) involved in dialogue activities.

   A user study was performed focusing on the exemplification of the computational framework proposed in this paper. Three selected dialogue activities were analysed from the developed scenarios in papers I and II, embedding conflict of intentions, the three strategies, and participant data.

   (a) The analysis showed that participants recognised when the conflict occurred and differentiated between *aligning* and *rejecting* strategies. The participants commented on the embodied agent's aligning strategy as cooperative, caring, soft, appropriately situated, empathetic, and positive. They also perceived the human actor's applied aligning strategy as considering the embodied agent's presence. It was observed that both human and embodied agents applied the *rejecting* strategy in the scenarios. The humans were observed being inattentive during dialogue activities and asked the embodied agent to leave them alone. There were instances when the embodied agent ignored the human's requests and intended topics. The embodied agent was also perceived as applying a rejecting strategy when it persisted and pursued its internal intentions.

   (b) The *partially aligning* strategy was considered natural to dialogue activities. This finding was intriguing as it indicates that partially aligning could potentially be considered a *good enough* level of agreement among the actors in a dialogue-based joint activity.

This paper addresses the research questions **RQ3** and **RQ4** on recognising and managing breakdown situations. Interesting to explore further will be the partially aligning strategy in user studies.

6.6 Paper VI

Maitreyee Tewari and Michele Persiani. Towards We-intentional Human-Robot Interaction using Theory of Mind and Hierarchical Task Networks. *In Proceedings of the 5th International Conference on Computer-Human Interaction Research and Applications - Humanoid, SciTePress, Pp. 291-299, 2021.*

Paper VI aimed to develop computational mechanisms allowing agents to plan and facilitate collaboration with other actors, including humans when a conflict of intention occurs.

This paper builds upon and contributes towards the following design implications in Paper I: *(1) co-constructing knowledge and learning, (2) co-constructing understanding and creating dynamic memories* about reasons for breakdown situations and applying strategies to manage those, and *(4) transparent decision making*.

Paper VI makes the following major contributions:

1. A novel formulation of Tuomela's *We-mode* and *I-mode* using hierarchical task network (HTN) planning and BDI-based Theory of Mind (ToM) to enable collaboration and planning under a conflict of intention. We assume that actors are in We-intention whenever they have to perform a joint activity. Before the agent can plan its joint actions, it needs to determine whether other actor/s have a conflict of intention or are co-creating one. We frame this conflict of intention as not something requiring management, which we explored in Paper V. Instead, the agent maps the conflict as the other actor operating in We-intention I-mode. In the absence of a conflict and given the agent is certain about the external intention, the agent assumes the other actor is operating in We-intention We-mode.

2. A computational framework embedding the We-mode and I-mode reasoning as follows: (i) initialisation of the ToM configuration, where the embodied agent represents itself and the human actor as equivalent probabilistic BDI frameworks; (ii) gathering observation using a dialogue activity, where the agent extracts actions from dialogue acts. The observations determine if the internal and external intentions align or conflict, (iii) update the intention

and the mode in the ToM, and (iv) plan using HTN. The agent's plan aligns with human expectations when it detects a We-intention We-mode. When the We-intention is in I-mode, the agent plans to achieve the intention efficiently.

The paper contributes to answering the research questions **RQ3** and **RQ4.** In the future, the proposed computational framework will be implemented and evaluated in user studies.

6.7 Towards a We-intentional Cognitive Architecture

This section describes an initial formulation of a cognitive architecture based on the design implications in papers I and II. The architecture comprises components necessary to manage dialogue activities, including breakdown situations (Figure 6.1). We designed the following components to be part of the cognitive architecture, commonly included in other architectures [45]: (1) *sensing and acting*, (2) *knowledge base*, (3) *situation assessment*, (4) *working memory*, (5) *reasoning*, (6) *planning/strategy selection*, and (7) *self-reflection.* Here, we also describe how the research contributions of this thesis fit in the cognitive architecture and indicate some of the future directions.

6.7.1 Sensing and Acting

To interact coherently, an agent needs components to perform sensing and acting. How an agent will act depends on its sensing capabilities. Information from the context and the environment can be captured using different kinds of sensors, for example, voice sensors to recognise speech and cameras for recognising human actors, their gestures, emotions, mood, movement, and objects in the environment.

An embodied agent can interact using body movement, positioning and orientation, gestures, and speech. Furthermore, acting in social scenarios with human actors requires an agent to behave appropriately by being empathetic, persuasive, or formal, for which deliberate behaviour selection can result from a reasoning module.

The design implications on the theme *understanding* in papers I and II suggest that the embodied agent must be able to sense human speech, gestures, expressions, and tone. To perform a dialogue activity with a human actor, the design implication on *interacting* in Paper II guides the embodied agent to be able to organise its body and be perceived as attentive and involved with the human actor. As indicated in papers I and II, the embodied agent must behave sympathetically and show emotions. Furthermore, the agent should provide information only as much as is

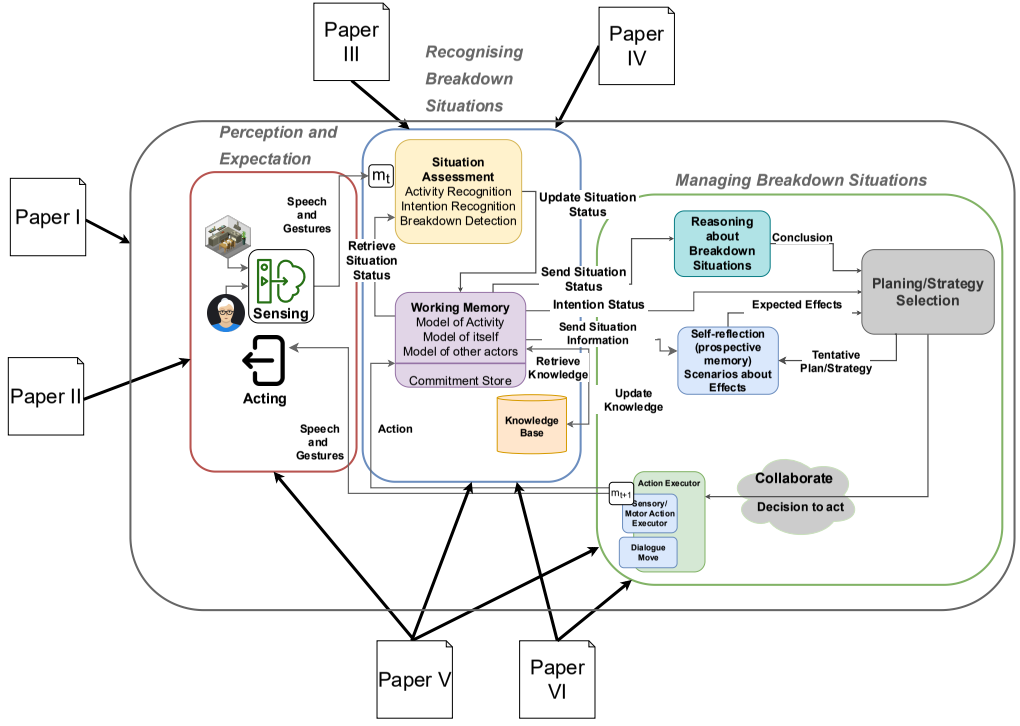


Figure 6.1: A cognitive architecture for an embodied agent to conduct dialogue activities with human actors. Themes and design implications in Paper I specify components for the cognitive architecture, indicated by a grey box. The insights from user studies (papers I and II) informed part of the cognitive architecture in the red box. Those enclosed with blue and green boxes are informed by the computational frameworks for recognising and managing breakdown situations (Paper III-VI).

required. The framework developed in Paper III, a scoring method to determine the informativeness and syntactic cohesion of dialogue turns, could be used by agents to produce optimal responses.

In summary, this interaction layer allows humans to infer the agent's intention and experience their behaviour. In terms of the activity-theoretical hierarchy of human activity, this layer manifests the operational level of activity, affecting how the agent is often subconsciously being experienced by other actors. This layer is further involved in mediating higher, more complex activity levels.

6.7.2 Knowledge Base

Guided by the design implication *2. co-constructing knowledge and learning* in Paper I, the knowledge base is a repository of knowledge that an agent can use, modify, and update to improve its understanding of itself, the other actors, dialogue activity and decision-making.

A potential knowledge base could be based on applying Engeström's Activity Model (EAM) [22] to a situation to instantiate generic knowledge relevant to a dialogue activity. The generic knowledge about the activity is combined with the knowledge about relevant human actors. The embodied agent builds models of the participating human and software actors, including their beliefs about themselves. Figure 3.1 illustrates how the EAM can represent generic knowledge about an activity, integrating contextual, social expertise as part of the embodied agent's beliefs makes this knowledge accessible and communicable when managing breakdown situations.

6.7.3 Working Memory

The working memory retrieves the information required for reasoning and planning. The Situation Assessment module, which interprets and exchanges information incoming from the sensing component, updates the working memory with the status of the situation. The status is about the ongoing activity, the internal intention of embodied agents and intentions of other actors, and if there is a breakdown situation. Working memory also retrieves relevant knowledge about who does what, social norms, and beliefs of self and other actors from the knowledge base. It further has a commitment store with recent actions performed by the actors as part of the ongoing dialogue activity.

6.7.4 Situation Assessment

The *Situation Assessment (SA) module* defines the process of creating an understanding of the ongoing situation. SA builds understanding based on the information from sensing the environment and information retrieved from the working memory. The output is the assessment of whether (i) the situation follows an ongoing activity (i.e., a case of maintaining we-intention), (ii) if there is no ongoing activity and a new activity is initiated (i.e., a case of transforming an I-intention into a we-intention), or (iii) if there is a breakdown situation and in this case what could be the reason for the breakdown.

Building upon the design implications for creating *understanding* in Paper I and Paper II, research presented in papers III-VI develop computational frameworks for different purposes to enable SA in an embodied agent.

In Paper III, a scoring method to calculate informativeness and cohesion could be used for SA. This scoring could enable SA to predict breakdown situations due to inconsistent or insufficient knowledge about facts, norms, procedures, intentions, and emotions or conflict of intentions (Section 6.3.)

In Paper IV, an annotation schema was developed and evaluated, enabling SA to capture several turns with long-dependency features. This could allow detection and reasoning of breakdown situations (Section 6.4.)

In Paper V, a computational framework driven by ASP-based reasoning for recognising and managing conflict of intentions was proposed. The SA module is responsible for inferring external intention (Section 6.5).

Paper VI proposed a computational framework exploiting probabilistic reasoning to determine and plan under conflict of intentions. The SA module is responsible for finding the probability distribution of the external intention.

6.7.5 Reasoning

The reasoning component decides if there is a breakdown situation and what causes it. The reasoning component uses the information from working memory to make a decision. The outcome is a decision.

Paper V provides mechanisms to reason under uncertain situations during joint activities, specifically a conflict of intention. We applied ASP to design mechanisms for agents to accept, reject, or partially accept a conflicting external intention. The output is a joint intention or a repaired intention.

In Paper VI, the framework reasons a conflict by measuring the distance between the probability distributions of the BDI model of the agent to that of the other actor/s. The output is the most likely goal and a plan.

Both papers V and VI attempt to realise a Theory-of-Mind (ToM) reasoning using BDI frameworks in the embodied agent.

6.7.6 Planning and Strategy Selection

The planning and strategy selection module determines the sets of actions the agent will perform. The module plans or selects the strategy based on the reasoning module's output.

In Paper V, the embodied agent has three strategies for planning and acting with other actors when a conflict of intention occurs. When rejecting the external intention, the embodied agent would display an *avoidant* strategy. When the agent accepts the conflicting intention, it applies an *agreeing* strategy. When the agent accepts only part of the conflicting intentions, it applies a *partially-agreeing* strategy. After the conflict has been repaired, the agent starts planning its actions.

Paper VI explored an alternative approach to managing conflict of intentions. Where instead of managing the conflict, the agent adapts its plan and aligns with human intentions as a strategy. The planner manages conflict of intentions by *performing tasks* under *constraints*. When there is no conflict due to other actors' We-mode We-intention, the planner imposes human expectations as constraints on tasks. In contrast, a conflict of intention occurs due to other actor/s being in I-mode; the planner imposes optimality constraints on the tasks.

6.7.7 Self-Reflection

The self-reflection component creates future scenarios based on the selected plan or dialogue strategy. Then, it evaluates the effects of the chosen course of action if they get effectuated in the human environment. The evaluation is based on incoming information about the assessed situation from the working memory. This information is compared with the chosen plan or strategy. Finally, the agent reflects on how it should act or not act in a given situation following social norms, division of tasks, and potential effects on the human actor's autonomy, integrity, and self-determination.

In papers I and II, an exemplification of self-reflection during an ongoing dialogue activity was provided in Scenario 1 (presented in Figure 3 in [86]). On detecting a conflict of intention, the embodied agent cooperates to align with the *human trail of thinking*. Adapting to human expectations was later defined as an *alignment* strategy in Paper V. Another potential behaviour after self-reflecting about conflict of intentions is to apply a *partially-aligning* strategy instead of *avoidant* strategy, displaying a partially-agreeing behaviour in Paper V. This partial agreement was found as good enough understanding for performing joint activities.

In papers I and II, the scenarios exemplify breakdown situations caused due to the absence of an optimal self-reflection module. For instance, Scenarios 3 and 4 (presented in Figures 5 and 6 in [86]) illustrate how the embodied agent apologises and stops the conversation when the person asks to be left alone. In this situation, the agent's self-reflection was not optimal. This could be stored as an *information* for the agent to reason about when it is socially appropriate to interrupt a person's activity.

Chapter 7

Contributions in the Perspectives of Human-centric AI

The human-centric AI perspective has been around for several decades, for example, in the early Human-Robot Interaction (HRI) research [26] and in the conceptualisation of Human Compatible AI [66]. More recently, human-centric AI has become pivotal in releasing the field of AI from the grasp of the *machine-centric* developments that have been driving AI research for the past couple of decades. For the human-centric AI perspective for healthcare, see [4].

This chapter reiterates the concerns about taking a pure machine-centric AI approach. We discuss how this thesis's research fits and contributes to the state-of-the-art of human-centric AI. Each of the six papers contributing to answering this thesis's research questions points towards their specific future directions. We conclude with an overarching and broader view of some of the limitations of this research and future work.

Machine-centric AI proliferated in the last two decades, improving the state-of-the-art in generating insights from big data, image/vision, and speech recognition. However, their integration into human society raises concerns due to experiences of being opaque, unfair, competitive to human intellectual capabilities, and decisions prone to be incorrect, such as in ChatGPT and other Large Language Models (LLMs) [52]. As an alternative and response to the machine-centric perspective, a perspective on human-centric AI has been conceptualised and developed to change the

focus of AI from being primarily efficiency-driven to that capable of collaborating, supporting, and enhancing human capabilities [61].

Steels characterises human-centric AI as focusing on developing agents with an understanding similar to humans. To develop the understanding, the agent should be aware of and adapt to the goals and intentions of the human and be able to explain their decisions. The agent should be capable of learning new knowledge by taking advice from human actors, and the ability to communicate and reflect on past interactions is considered important. Self-reflection allows an agent to consider ethical and moral standards, identify and manage failures, and support human actors in their activities [80].

Furthermore, Steels emphasises that a human-centred understanding is a composition of meaning formation at multiple levels. These levels of meaning beginning with low complexity are *formal, factual, expressional, social, conventional,* and *intrinsic* at the highest level of complexity. Meaning at the lowest (formal) level corresponds to understanding the standard properties derived from perception. The factual level contains the facts about actors, entities, and roles, similar to human semantic memory. The expressional level creates meaning about the involved actors' intentions, goals, interests, and motivations. The social level is about social relations and how each agent performs their activities within a community [80]. The three levels: factual, expressional, and social levels correspond to the *operational* and *action* levels in Activity theory (Figure 7.1) [42]. The conventional level represents the norms depicting the historical and cultural context, corresponding to the activity level in Activity theory. The highest intrinsic level is to understand the inherent reason behind the behaviour of an agent. In activity-theoretical terms, this level corresponds to the human's motives and needs, which the human may be unaware of [42].

Recently, Crowley and colleagues [16] provided a hierarchical framework for collaborative AI as a research agenda motivated by the human-centric perspective. The hierarchical framework integrates agents with abilities to learn, comprehend, and explain when collaborating with human actors. The hierarchy of collaboration consists of (1) a *reactive* level, where actions of one agent are sensed by the other, triggering an action in response (corresponding to a large part of the robotics research, e.g., navigation, affective responses, error management); (2) a *situational* level, where actions are informed by the shared understanding of gathered evidence and predicted consequences of the situation consisting of factual information (challenges addressed by contemporary research). Next is (3) the *operational* level, the agent plans and executes tasks to achieve a desired goal based on information about goals, tasks, plans, and actions of other agents (challenges addressed by contemporary research); (4) The *praxical* level of collaboration requires an agent to adhere

to task protocols and to perform interactions with human actors following social and conventional norms (introducing the challenges in the research community for example in social robotics). The highest, (5) *creative* level improves and builds upon each other's knowledge and ideas by understanding the other participant (defined to be beyond the current state-of-the-art).

Activity theory was foundational in this thesis to study, formalise, and implement the human-centric design of social and intelligent embodied agents. Activity theory provided a model of *human collective activities* embedding components at different levels of complexity. Activity theory also provided notions of *contradictions* (referred to as breakdown situations in this thesis) leading to a *transformation* between these levels and from one central activity to alternative activities (Chapter 2). Similar models recently suggested by AI researchers as summarised in this chapter, highlight the need for the agents to understand and collaborate with human actors at different levels of complexity during an activity [16, 80]. The contributions of this thesis were mapped to the models to elaborate how this work spans different levels of human activity and human-agent collaboration (Figure 7.1).

This thesis contributes to human-centric AI research by providing increased knowledge and understanding of how people perceive and experience breakdown situations across the levels of activity. The thesis additionally provides design implications addressing aspects (understanding, interacting, relating) corresponding to the different levels of dialogue activity. Embedded in the qualitative studies in papers I and II was the question on the ethical aspects around placing the embodied agent in the home. It was expected from the embodied agents cohabiting the human social space to know the *norms of the house* and is secure to use. The ethical considerations of the studies conducted in this thesis and the implications for embodied agents are further discussed in Section 7.1.

The proposed cognitive architecture conceptualised based on the design implications in papers I and II, illustrates how different mechanisms of human-centric AI embed awareness and reasoning of human activity, intentions, and associated breakdown situations. Planning methods that adapt to human preferences communicating using multi-modal dialogues, and reflecting on past interactions to improve and manage breakdown situations were provided. Strategies and behaviour for embodied agents guided by design implications, such as following the human's trail of thinking, were proposed. Such strategies can manage conflicting intentions under specific conditions or build new knowledge in consultation with the human actor. This contribution aligns with the objective of human-centric AI.

The developed understanding of breakdown situations when embodied agents engage in dialogue activities with human actors contributed to the development of formal frameworks for managing meaning formation at formal, factual, expres-

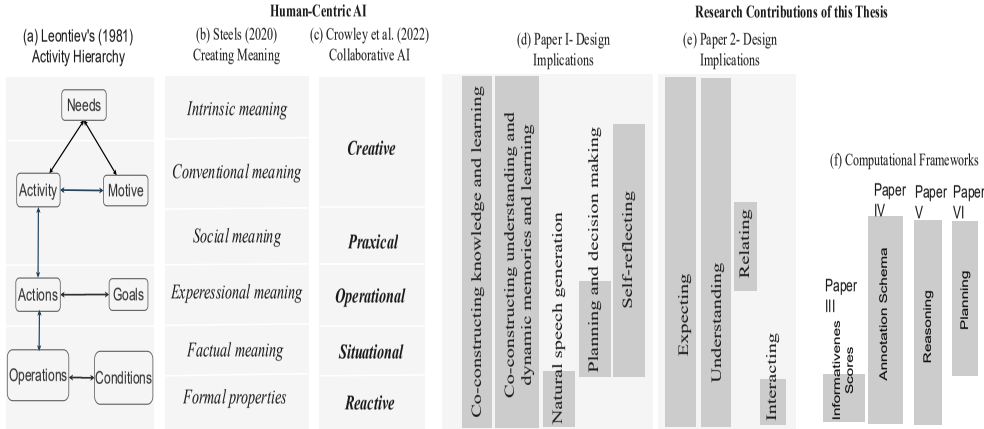


Figure 7.1: An overview of the levels of complexity of human activities involving embodied agents as conceptualised by Activity theory [48] and Human-centric AI perspective [16, 80]. On the right is illustrated how the contributions of this thesis address different levels of complexity in human activity.

sional, and social levels (Papers III-VI, Figure 7.1) [80]. These levels for dialogue activities can be interpreted as the syntax, semantics, knowledge, emotion, social, and intention levels embedded in the design implications in Paper II and partially in the annotation schema in Paper IV.

Crowley and colleagues exemplify their hierarchical framework of collaboration through hypothetical scenarios [16]. Evidence that supports their framework specific to daily living scenarios is provided in this work. The contributions, such as themes, taxonomy, and computational frameworks, provided a hybrid approach complementing and supporting Crowley and colleagues [16] learning-based methods for understanding.

The illustrated perspectives on human-centric AI as the complex multi-faceted domain provided by Steels, Crowley and colleagues are explained by the activity-theoretical models of the human.

7.1 Ethical Considerations and Implications

Studies in papers I, II, III, and V involved human subjects. Thus, ethical principles and regulations were followed with respect to the study setup, methodology, and the resulting frameworks. Papers I and II also explored the ethical aspects relating to embodied agents having dialogues with people about healthcare and well-being. We discuss both ethical aspects in the subsequent section.

To ensure compliance with ethical principles and regulations, the researchers discussed and reflected upon the need for ethical approval. The primary researcher was in contact with the university's legal officer, inquiring about the need for ethical approval. It was concluded that ethical approval was not needed for the following reasons: (1) it was sufficient to register the study in the University registry, (2) provide the information in the consent form to the participants about when their data is being processed and (3) there was no direct interaction between participants and the embodied agent. To comply with the General Data Protection Regulations (GDPR), the legal officer at the University recommended we use "Task in the Public Interest" as the lawful ground for processing personal data. Two separate debriefing and consent documents were prepared, aimed towards the study's volunteers and participants. GDPR was emphasised in the consent and debriefing documents, explaining the study's objective and the rights of the volunteers and participants. Following the university guidelines for studies processing people's personal data, the studies in papers I, II, and V have been registered on the university's registry. Participants provided consent for the study in Paper III and anonymously answered the survey online; no personal data was gathered.

For studies in papers I, II, and V, minimum personally identifiable information, including the email, gender and age of the participants, was recorded and stored locally on the primary researcher's computer. The email was used to communicate with the participants and assign unique identifiers. This was required so that if a participant wanted to withdraw from the study, then we could delete their data. Age and gender were used to assign the participants to older, middle-aged, and young adult groups for analysis purposes. After the analysis, all the recordings were deleted. Only the anonymised interview transcriptions and an Excel sheet with email addresses and unique identifiers were retained locally on the primary researcher's computer.

7.1.1 Ethics Considerations Regarding the Domain of Study

Our discussions in papers I, II, and V relating to health and well-being dialogues between humans and embodied agents raise the following main ethical concerns: (1) the ethical concern of embodied agents replacing caregivers, which has also been raised in research community [99], and (2) roles, obligations and behaviours of embodied agents in human social spaces.

To address concern 1 from a design and engineering perspective, the scenarios based on Activity Theory provided us with the concepts to design the embodied agent conforming to the division of tasks for joint activities. This allowed situating the embodied agent as one of the members contributing to a joint activity in which the caregivers and the person being cared for perform their roles. Furthermore, as a we-intentional actor in joint activities, the embodied agent was designed to adopt the motives of the human being cared for to address their needs. This makes the agent a member contributing to a joint activity rather than something with the potential to substitute for another community member. Our studies showed that depending on the embodied agent's actions, participants perceived the agent embodying different roles of an assistant, companion, and caregiver when reminding about medication and pursuing the volunteer to seek medical assistance for back pain. Thus, participants, differing views on the agent's role were captured during our studies.

Regarding concern 2, the qualitative methodology with semi-structured interviews allowed participants to reflect on ethical aspects, social norms, expected roles, behaviours, and obligations of embodied agents. The studies included participants from different nationalities and broad age groups. This allowed us to capture various perspectives, contributing to insights on ethical concerns. The perspectives involved deploying security and privacy measures consented functionalities of embodied agents, acceptable behaviour, design, roles, and social norms.

There was an agreement between older and younger participants about the capabilities to manage emergencies. Older and younger participants focused on different ethical concerns relating to embodied agents. The difference in focus on what ethical considerations need to be embedded could be attributed to the familiarity with technology. In the activity theoretical sense, when exposed to something new, people have to focus on learning the operations. This could be why older people focused more on privacy, security, and understanding how the embodied agent functions. With more familiarity, the focus moves from lower operational levels to higher levels of activity. For instance, younger people focused on how the agent should behave and the social norms it should conform to in dialogues with people.

In summary, our methodology contributed to designing embodied agents as a member of a community contributing to managing the health and well-being of a human needing care. Instead of being a substitute of caregivers, our embodied agent supports them to achieve their goals. The studies highlight what people consider essential ethical aspects that should be embedded in socially intelligent embodied agents managing health and well-being dialogue activities with humans.

7.2 Limitations and Future Work

The human-centric AI perspective motivated us to investigate dialogue activities in natural settings, building on well-established theories on human activity and their reasoning. This highlighted the complexity of dialogue activities due to the co-creation of fragmented understanding leading to breakdown situations.

Taking a human-centric AI perspective and applying theories from psychology and social sciences were necessary to fulfil the aim of this work. Activity theory provided an alternative perspective on representing breakdown situations as *opportunities to learn something new*, instead of the traditional view of treating them as failures or errors [39, 57, 77]. This approach introduced novel perspectives on the situations we were studying and expanded the scope of our research.

This thesis's research scope spanned different levels following Leontiev's *hierarchy of activity* [47, 48]. This added complexity to the tasks related to developing formal representations and computational frameworks for managing breakdown situations.

The research findings in papers I and II are in line and encompass the various requirements from robot companions as defined by Ahmed and colleagues [2]. These requirements span embodied agents' physical, social, emotional, and safe interaction abilities, which map to the interacting and relating factors in Paper II. Additionally, a detailed account of the understanding aspect is provided in Paper

II, which is central to any interaction between human and embodied agents and was mostly overlooked (the exception being emotion recognition) in the survey on companion embodied agents [2]. This allows the aspects to be used as research agendas for designing and building embodied agents intended to interact with people.

Taking the *relating* aspect as a research agenda, immediate future work is to explore how *agency* relating to the agent's varying roles in tasks involving dialogues between humans and agents connects to the factors in understanding, relating, and interacting. Furthermore, the actors interacting with an agent in various roles will be used to explore how breakdown situations are perceived and acted upon in those roles.

The research findings of the thesis on the state-of-the-art relating to embodied agents interacting using dialogues with humans being merely command-driven aligns with the findings in Ahmed and colleagues' survey [2]. The Activity Theory-based analysis of scenarios resulting in dialogues as a co-creation exemplifies and supports the claims about *co-performance* [46] necessary for companionship between embodied agents and humans [2]. Additionally, strategies to align, not align, and partially align in Paper V are the computational mechanisms that can enable co-performance with embodied agents. A preliminary evaluation of these mechanisms is presented in Paper V, where people could discriminate when the actors aligned or not aligned their intentions. Interesting to explore further would be partially aligning strategy as it was considered a good enough agreement for potential joint activity.

The two computational frameworks presented in papers III and IV delve into building understanding mechanisms for agents to embed meaning at lower levels of syntax, moving up to the semantics and then to the higher levels of social rules and policies to organise dialogues. This is different from the current state-of-the-art large language models (LLMs) [100], which creates meaning only related to syntax [6]. The theories underlying and embedded in these frameworks ranged from socio-linguistics [76, 81] explaining how people organise dialogue activities, philosophy [32], and computational linguistics [10], explaining the different types of meaning people create in dialogues. This aligns with the kind of meaning formation, as argued to be necessary for creating understanding in agents similar to human understanding, by Bender and colleagues [6]. However, this work needs further evaluation. An interesting opportunity for future research would be to extend the frameworks with the knowledge about the context embedding the related breakdown situations. Another future work could be integrating and building context in a dialogue activity by applying theory-of-mind-like reasoning, explored in papers V and VI.

Crowley and colleagues define the ability to learn, understand, and explain as essential for embodied agents to collaborate with human actors [16]. This thesis deepens the *understanding* of breakdown situations embedded in dialogue activities. This focus paves the way for future research to investigate and design capabilities for embodied agents to *learn* to adapt and *explain* their decisions to humans. These capabilities could be integrated into our proposed cognitive architecture (Chapter 6). Future work would further evaluate, aggregate, and connect computational frameworks proposed in this thesis within the cognitive architecture to manage dialogue activities and related breakdown situations between humans and socially intelligent agents.

References

[1] John Aberdeen and Lisa Ferro. “Dialogue patterns and misunderstandings”. In: *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*. Switzerland: ISCA, 2003, pp. 1–5.

[2] Eshtiak Ahmed, Oğuz Oz. Buruk, and Juho Hamari. “Robots as Human Companions: A Review”. In: *Pacific Asia Conference on Information Systems*. Association for Information Systems. 2022, p. 246.

[3] Davide Ancona and Viviana Mascardi. “Coo-BDI: Extending the BDI Model with Cooperativity”. In: *Declarative Agent Languages and Technologies*. Berlin, Heidelberg: Springer, 2004, pp. 109–134.

[4] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Enrico Coiera, and Yvonne Rogers. “Introduction to the Special Issue on Human-Centred AI in Healthcare: Challenges Appearing in the Wild”. In: *ACM Transaction on Computer-Human Interaction* 30.2 (2023), p. 12.

[5] Chitta Baral. “Declarative programming in AnsProlog- introduction and preliminaries”. In: *Knowledge representation, reasoning and declarative problem solving with Answer sets*. 2001. Chap. 1, p. 10.

[6] Emily M. Bender and Alexander Koller. “Climbing towards NLU: On meaning, form, and understanding in the age of data”. In: *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*. 2020, pp. 5185–5198.

[7] Susanne Bødker and Peter B. Andersen. “Complex mediation”. In: *Human-Computer Interaction* 20.4 (2005), pp. 353–402.

[8] Michael E. Bratman. *Intention, plans, and practical reason*. Harvard University Press, 1987.

[9] Michael E. Bratman. “Shared Intention”. In: *Ethics* 104.1 (1993), pp. 97–113.

[10] Harry Bunt. “Dynamic interpretation and dialogue theory”. In: *The structure of multimodal dialogue* 2 (1999), pp. 139–166.

[11] Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. “Towards an ISO Standard for Dialogue Act Annotation”. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. Valletta, Malta: European Language Resources Association (ELRA), May 2010.

[12] Harry Bunt, Volha Petukhova, Andrei Malchanau, Alex Fang, and Kars Wijnhoven. “The DialogBank: dialogues with interoperable annotations”. In: *Language Resources and Evaluation* 53.2 (2019), pp. 213–249.

[13] Harry Bunt, Volha Petukhova, Andrei Malchanau, Kars Wijnhoven, and Alex Fang. “The DialogBank”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Portorož, Slovenia: European Language Resources Association (ELRA), 2016, pp. 3151–3158.

[14] Victoria Clarke, Virginia Braun, and Nikki Hayfield. “Thematic analysis”. In: *Qualitative psychology: A practical guide to research methods* 3 (2015), pp. 222–248.

[15] Philip R. Cohen, Hector J. Levesque, and Ira A. Smith. “On team formation”. In: *Synthese Library* (1997), pp. 87–114.

[16] James L. Crowley, Joëlle Coutaz, Jasmin Grosinger, Javier Vazquez-Salceda, Cecilio Angulo, Alberto Sanfeliu, Luca Iocchi, and Anthony G. Cohn. “A Hierarchical Framework for Collaborative Artificial Intelligence”. In: *IEEE Pervasive Computing* (2022), pp. 1–10.

[17] María De Los Ángeles Gómez González. “Cohesion”. In: *The International Encyclopedia of Language and Social Interaction*. American Cancer Society, 2015, pp. 1–12.

[18] Frank Dignum, David N. Morley, Elizabeth A. Sonenberg, and Lawerance Cavedon. “Towards socially sophisticated BDI agents”. In: *Proceedings of the Fourth International Conference on MultiAgent Systems*. 2000, pp. 111–118.

[19] Mateusz Dolata, Dzmitry Katsiuba, Natalie Wellnhammer, and Gerhard Schwabe. “Learning with Digital Agents: An Analysis based on the Activity Theory”. In: *Journal of Management Information Systems* 40.1 (2023), pp. 56–95.

[20] Phan Minh Dung. “On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games”. In: *Artificial Intelligence* 77.2 (1995), pp. 321–357.

[21] Alessandro Duranti and Charles Goodwin. “Rethinking context: an Introduction”. In: *Rethinking Context: Language as an Interactive Phenomenon*. 11. Cambridge University Press, 1992.

[22] Yrjö Engeström. “Activity theory and individual and social transformation”. In: *Perspectives on activity theory*. Vol. 19. 38. 1999, pp. 19–38.

[23] Yrjö Engeström. “Expansive Visibilization of Work: An Activity-Theoretical Perspective”. In: *Computer Supported Cooperative Work (CSCW)* 8.1-2 (1999), pp. 63–93.

[24] Yrjö Engeström. “Objects, contradictions and collaboration in medical cognition: an activity-theoretical perspective.” In: *Artificial Intelligence in Medicine* 7.5 (1995), pp. 395–412.

[25] Yrjö Engeström. “The Emergence of Learning Activity as a Historical Form of Human Learning”. In: *Learning by Expanding: An Activity-Theoretical Approach to Developmental Research*. Cambridge University Press, 2019, pp. 25–108.

[26] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. “A survey of socially interactive robots”. In: *Robotics and Autonomous Systems* 42.3 (2003), pp. 143–166.

[27] Ilche Georgievski and Marco Aiello. “An Overview of Hierarchical Task Network Planning”. In: arXiv, 2014.

[28] Spiros Georgiladakis, Georgia Athanasopoulou, Raveesh Meena, José Lopes, Arodami Chorianopoulou, Elisavet Palogiannidi, Elias Iosif, Gabriel Skantze, and Alexandros Potamianos. “Root Cause Analysis of Miscommunication Hotspots in Spoken Dialogue Systems.” In: *inInterspeech*. ISCA, 2016, pp. 1156–1160.

[29] Barbara Gonsior, Christian Landsiedel, Antonia Glaser, Dirk Wollherr, and Martin Buss. “Dialog strategies for handling miscommunication in task-related HRI”. In: *2011 RO-MAN*. IEEE, 2011, pp. 369–375.

[30] Robert M. Gordon. “Folk Psychology as Simulation”. In: *Mind & Language* 1.2 (1986), pp. 158–171.

[31] Paul Green and Lisa Wei-Haas. “The rapid development of user interfaces: Experience with the Wizard of Oz method”. In: *Proceedings of the Human Factors Society Annual Meeting*. Vol. 29. 5. SAGE Publications, 1985, pp. 470–474.

[32] Herbert P. Grice. “Logic and conversation”. In: *Syntax and Semantics 3: Speech Acts* (1975), pp. 41–58.

[33] Esteban Guerrero, Ming-Hsin Lu, Hsiu-Ping Yueh, and Helena Lindgren. “Designing and evaluating an intelligent augmented reality system for assisting older adults’ medication management”. In: *Cognitive Systems Research* 58 (2019), pp. 278–291.

[34] Esteban Guerrero, Juan Carlos Nieves, and Helena Lindgren. “An activity-centric argumentation framework for assistive technology aimed at improving health”. In: *Argument & Computation* 7.1 (2016), pp. 5–33.

[35] Esteban Guerrero, Maitreyee Tewari, Panu Kalmi, and Helena Lindgren. “Forming We-intentions under breakdown situations in human-robot interactions”. In: *Computer Methods and Programs in Biomedicine* 242 (2023), pp. 107–817.

[36] Charles L. Hamblin. “Fallacies”. In: *Tijdschrift Voor Filosofie* 33.1 (1970), pp. 183–188.

[37] Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. “Towards Taxonomy of Errors in Chat-oriented Dialogue Systems”. In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague, Czech Republic: ACL, 2015, pp. 87–95.

[38] Graeme Hirst, Susan McRoy, Peter Heeman, Philip Edmonds, and Diane Horton. “Repairing conversational misunderstandings and non-understandings”. In: *Speech communication* 15.3-4 (1994), pp. 213–229.

[39] Shanee Honig and Tal Oron-Gilad. “Understanding and resolving failures in human-robot interaction: Literature review and model development”. In: *Frontiers in Psychology* 9 (2018), p. 21.

[40] Chien-Ming Huang and Bilge Mutlu. “Robot behavior toolkit: Generating effective social behaviors for robots”. In: *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2012, pp. 25–32.

[41] Ian Hutchby and Robin Wooffitt. *Conversation Analysis*. Wiley, 2008.

[42] Victor Kaptelinin and Bonnie A. Nardi. *Acting with technology: Activity theory and interaction design*. MIT press, 2006.

[43] Victor Kaptelinin and Bonnie A. Nardi. “Activity Theory in HCI: Fundamentals and Reflections”. In: *Synthesis Lectures on Human-Centered Informatics* 5 (2012).

[44] David Kinny, Elizabeth Sonenberg, Magnus Ljungberg, Gil Tidhar, Anand S. Rao, and Eric Werner. “Planned team activity”. In: *European Workshop on Modelling Autonomous Agents in a Multi-Agent World*. Springer, 1994, pp. 227–256.

[45] Iuliia Kotseruba, Oscar J. Avella Gonzalez, and John K. Tsotsos. “A Review of 40 Years of Cognitive Architecture Research: Focus on Perception, Attention, Learning and Applications”. In: *Clinical Orthopaedics and Related Research* abs/1610.08602 (2016).

[46] Lenneke Kuijer and Elisa Giaccardi. “Co-performance: Conceptualizing the role of artificial agency in the design of everyday life”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, pp. 1–13.

[47] Aleksei N. Leontiev. *Activity, consciousness, and personality.* Moscow, Russia: Prentice-Hall, 1978.

[48] Aleksei N. Leontiev. *Problems of the development of the mind.* Moscow, Russia, 1981.

[49] Xu Liang, Ruo Wang, and Guohua Bai. “A Multi-Agent System Based on Activity Theory for Collaborative Network Learning”. In: *2009 First International Workshop on Education Technology and Computer Science*. Vol. 1. 2009, pp. 392–397.

[50] Per Linell. “Troubles with Mutualities: Towards a Dialogical Theory of Misunderstanding and Miscommunication”. In: *Mutualities in Dialogue*. UK: Cambridge University Press, 1995. Chap. 8, pp. 176–212.

[51] Ramón López-Cózar, Zoraida Callejas, and David Griol. “Using knowledge of misunderstandings to increase the robustness of spoken dialogue systems”. In: *Knowledge-Based Systems* 23.5 (2010), pp. 471–485.

[52] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. *When Not to Trust Language Models: Investigating Effectiveness and Limitations of Parametric and Non-Parametric Memories*. 2022. arXiv: 2212.10511 [cs.CL].

[53] Matthew Marge and Alexander I. Rudnicky. “Comparing Spoken Language Route Instructions for Robots across Environment Representations”. In: *Proceedings of the SIGDIAL 2010 Conference*. Tokyo, Japan: Association for Computational Linguistics, Sept. 2010, pp. 157–164.

[54] Matthew Marge and Alexander I. Rudnicky. “Miscommunication detection and recovery in situated human-robot dialogue”. In: *ACM Transactions on Interactive Intelligent Systems* 9.1 (2019).

[55] Michael McTear. “Conversation modelling for chatbots: current approaches and future directions”. In: *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung* (2018), pp. 175–185.

[56] Michael McTear. “Spoken dialogue technology: toward the conversational user interface”. In: Springer Science and Business Media, 2004, pp. 113–116.

[57] Raveesh Meena, José Lopes, Gabriel Skantze, and Joakim Gustafson. “Detection of Miscommunication in Spoken Dialogue Systems”. In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague, Czech Republic: ACL, Sept. 2015, pp. 354–363.

[58] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. “To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot”. In: *Frontiers in Robotics and AI* 4 (2017), p. 21.

[59] Erik T. Mueller. “Chapter 15 - Commonsense Reasoning Using Answer Set Programming”. In: *Commonsense Reasoning (Second Edition)*. Second Edition. Boston: Morgan Kaufmann, 2015, pp. 249–269.

[60] Joakim Nivre. *Dependency grammar and dependency parsing*. 1959. 2005, pp. 1–32.

[61] Andrzej Nowak, Paul Lukowicz, and Pawel Horodecki. “Assessing Artificial Intelligence for Humanity: Will AI be Our Biggest Ever Advance ? or the Biggest Threat [Opinion]”. In: *IEEE Technology and Society Magazine* 37.4 (2018), pp. 26–34.

[62] Michele Persiani and Maitreyee Tewari. “Mediating Joint Intention with a Dialogue Management System”. In: *1st International Workshop on New Foundations for Human-Centered AI*. RWTH Aachen University. 2020, pp. 79–82.

[63] Marco Ragni, Andrey Rudenko, Barbara Kuhnert, and Kai O. Arras. “Errare humanum est: Erroneous robots in human-robot interaction”. In: *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2016, pp. 501–506.

[64] Anand S. Rao and Michael P. Georgeff. “BDI agents: from theory to practice”. In: *Proceedings of the First International Conference on Multiagent Systems (ICMAS)*. Vol. 95. MIT Press, 1995, pp. 312–319.

[65] Alessandro Ricci, Andrea Omicini, and Enrico Denti. “Activity Theory as a Framework for MAS Coordination”. In: *Engineering Societies in the Agents World III*. Springer, 2003, pp. 96–110.

[66] Stuart Russell. In: *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Random House LLC, 2019. Chap. 1. If We Succeed, pp. 1–12.

[67] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. “Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust”. In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 141–148.

[68] Emanuel A. Schegloff and Harvey Sacks. “Opening up closings”. In: *Semiotica* 8.4 (1973), pp. 289–327.

[69] Niels Schütte, John Kelleher, and Brian M. Namee. “Clarification dialogues for perception-based errors in situated human-computer dialogues”. In: *Proceedings of the 2014 ACM Workshop on Multimodal, Multi-Party, Real-World Human-Robot Interaction*. ACM, 2014, pp. 25–26.

[70] Niels Schütte, Brian M. Namee, and John Kelleher. “Robot perception errors and human resolution strategies in situated human–robot dialogue”. In: *Advanced Robotics* 31.5 (2017), pp. 243–257.

[71] Niels Schütte, Brian M. Namee, and John Kelleher. “Robot perception errors and human resolution strategies in situated human–robot dialogue”. In: *Advanced Robotics* 31.5 (2017), pp. 243–257.

[72] Wilfred Sellars. “Science and Metaphysics: Variations on Kantian Themes”. In: *Philosophy* 45.171 (1970), pp. 66–70.

[73] Wilfrid Sellars. “Empiricism and the Philosophy of Mind”. In: *Minnesota studies in the philosophy of science* 1.19 (1956), pp. 253–329.

[74] Wilfrid Sellars. “On Reasoning about Values”. In: *American Philosophical Quarterly*. Vol. 17. JSTOR, 1980, pp. 81–101.

[75] Sofia Serholt. “Breakdowns in children’s interactions with a robotic tutor: A longitudinal study”. In: *Computers in Human Behavior* 81 (Apr. 2018), pp. 250–264.

[76] Jack Sidnell and Tanya Stivers. *The handbook of conversation analysis*. Vol. 121. John Wiley & Sons, 2012.

[77] Gabriel Skantze. “Error Handling in Spoken Dialogue Systems - Managing Uncertainty, Grounding and Miscommunication”. PhD thesis. KTH Computer Science and Communication, 2007.

[78] Gabriel Skantze. “Galatea: A discourse modeller supporting concept-level error handling in spoken dialogue systems”. In: *Recent Trends in Discourse and Dialogue*. Lisbon, Portugal: Springer, 2008, pp. 155–189.

[79] Clay Spinuzzi. “Trying to predict the future: third-generation activity theory’s codesign orientation”. In: *Mind, Culture, and Activity* 27.1 (2020), pp. 4–18.

[80] Luc L. Steels. “Personal dynamic memories are necessary to deal with meaning and understanding in human-centric AI”. In: *NeHuAI@ECAI*. 2020.

[81] Tanya Stivers. “Sequence Organization”. In: *The Handbook of Conversation Analysis*. UK: Wiley-Blackwell, 2012. Chap. 10, pp. 191–209.

[82] Maitreyee Tewari. “Formalization of Dialogues from Movie Corpus using DAMSL Annotation Scheme as Cooperating Distributed Grammar Systems”. In: *Report UMINF* (2021).

[83] Maitreyee Tewari and Suna Bensch. “Natural language communication with social robots for assisted living”. In: *Robots for Assisted Living-IROS’2018 Workshop, IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2018, pp. 1–4.

[84] Maitreyee Tewari, Suna Bensch, Thomas Hellström, and Kai-Florian Richter. “Modelling Grice’s Maxim of Quantity as Informativeness for Short Text”. In: *ICLLL 2020: The 10th International Conference in Languages, Literature, and Linguistics*. 2020, pp. 1–7.

[85] Maitreyee Tewari, Monika Jingar, and Suna Bensch. “A Hybrid Model to Classify Sudden Topic Change, Misunderstanding and Non-understanding in Human Chat-bot Interaction”. Preprint manuscript on Diva at webpage `http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-174674`. 2020.

[86] Maitreyee Tewari and Helena Lindgren. “Expecting, understanding, relating, and interacting-older, middle-aged and younger adults’ perspectives on breakdown situations in human–robot dialogues”. In: *Frontiers in Robotics and AI* 9 (2022).

[87] Maitreyee Tewari and Helena Lindgren. “Younger and Older Adults’ Perceptions on Role, Behavior, Goal and Recovery Strategies for Managing Breakdown Situations in Human-Robot Dialogues”. In: *Proceedings of the 9th International Conference on Human-Agent Interaction*. ACM Digital Library, 2021, pp. 433–437.

[88] Maitreyee Tewari and Michele Persiani. “Towards We-intentional Human-Robot Interaction using Theory of Mind and Hierarchical Task Network”. In: *The 5th International Conference on Computer-Human Interaction Research and Applications - Humanoid*. SciTePress, 2021, pp. 291–299.

[89] Maitreyee Tewari and Michele Persiani. “Variational Autoencoding Dialogue Sub-Structures Using a Novel Hierarchical Annotation Schema”. In: *2020 6th IEEE Congress on Information Science and Technology (CiSt)*. IEEE. 2021, pp. 334–341.

[90] Michael Tomasello and Hannes Rakoczy. “What Makes Human Cognition Unique? From Individual to Shared to Collective Intentionality”. In: *Mind & Language* 18.2 (2003), pp. 121–147.

[91] Raimo Tuomela. “Joint Intention, We-Mode and I-Mode”. In: *Midwest Studies In Philosophy* 30.1 (2006), pp. 35–58.

[92] Raimo Tuomela. *The philosophy of social practices: A collective acceptance view*. Cambridge University Press, 2002.

[93] Raimo Tuomela and Kaarlo Miller. “We-Intentions”. In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 53.3 (1988), pp. 367–389.

[94] Christian V. Scheve and Sven Ismer. “Towards a theory of collective emotions”. In: *Emotion review* 5.4 (2013), pp. 406–413.

[95] Lev S. Vygotsky. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, 1978.

[96] Douglas Walton and Erik C.W. Krabbe. “Dialogues: Types, Goals and Shifts”. In: *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press, 1995. Chap. 3, pp. 65–117.

[97] Douglas Walton and Erik C.W. Krabbe. “Introduction”. In: *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press, 1995. Chap. 1, pp. 16–28.

[98] Michael Wooldridge, Simon Parsons, Gerhard Goos, Juris Hartmanis, and Jan V. Leeuwen. “Intention Reconsideration Reconsidered”. In: *Intelligent Agents V: Agents Theories, Architectures, and Languages*. Vol. 1555. Springer, 1999, pp. 63–79.

[99] Ricarda Wullenkord and Friederike Eyssel. “Societal and Ethical Issues in HRI”. In: *Current Robotics Reports* 1 (3 2020), pp. 2662–4087.

[100] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. *A Survey of Large Language Models*. 2023. arXiv: 2303.18223 [cs.CL].