Decision-theoretic, Personality-based Management of Turn-taking Conflicts in Multimodal Dialogue Systems

Doctoral Thesis

University of Augsburg
Faculty of Applied Computer Science
Chair of Human-Centered Artificial Intelligence

submitted by
Kathrin Janowski
in January 2024

Supervisor and First Reviewer:	Prof. Dr. Elisabeth André
Second Reviewer:	Prof. Dr.-Ing. Johannes Schilp
Date of Defense:	April 8^{th} 2024
First Examiner:	Prof. Dr. Elisabeth André
Second Examiner:	Prof. Dr.-Ing. Johannes Schilp
Third Examiner:	Prof. Dr. Birgit Lugrin

To my parents who raised me as a proud computer geek.

Abstract

Humans tend to interpret the behavior of robots and virtual characters in human terms. Therefore, interaction designers need to ensure that the agent's behaviors align with the personality that users are meant to associate with it. One such behavior is the turn-taking in conversations with the user. In particular, overlaps and interruptions are loaded with stereotypes about dominance, but also positive phenomena such as shared enthusiasm. Silence can be awkward or a sign of patient listening.

In order to generate consistent behavior for a wide range of conversational agents - from obedient home assistants to antagonists in training simulations - a psychologically sound model is required. This thesis therefore reviewed existing theories about how personality and interpersonal attitude are reflected in turn-taking behavior, specifically the timing of speech activity and the accompanying gaze signals. A decision-theoretic approach was then chosen to model idealized human-like reasoning and thereby strike a balance between generating more natural agent behavior and meeting the heightened expectations that humans have towards a rational machine.

The concept presented here consists of three parts. The *influence diagram* chooses the turn-taking behavior, such as starting to speak or averting the gaze, that best fulfills the agent's goals. The *Participant Framework* connects this behavior model to the dialogue manager, providing context information for the influence diagram's decisions and using said decision to regulate the dialogue flow. Finally, the *RobotEngine Framework* connects the dialogue application to different virtual and robotic agents in a way that keeps the turn-taking behavior separate from the agents' implementation.

Finally, the developed behavior model was tested in two different example setups. A non-interactive prototype with a simplified behavior model had two virtual characters talking to each other, implemented as separate processes and limited to explicit verbal communication. This limitation was meant to simulate the incomplete knowledge that a conversational agent could obtain from a human user. An online study confirmed that the generated behaviors

led to personality judgments in line with theory and related works. Afterward, an interactive prototype was set up with incremental speech recognition and gaze detection. A preliminary evaluation was conducted by analyzing recordings of human-agent conversations. While the generated behavior variations were in line with the expected patterns, the interactivity introduced numerous challenges that will require a more thorough analysis in the future. Nevertheless, important lessons were learned and summarized as recommendations for evaluating such a system.

Zusammenfassung

Menschen neigen dazu, das Verhalten von Robotern und virtuellen Figuren nach menschlichen Maßstäben zu interpretieren. Daher müssen Interaktionsdesigner sicherstellen, dass das Verhalten des Agenten mit der Persönlichkeit übereinstimmt, die die Benutzer mit ihm verbinden sollen. Eines dieser Verhalten ist die Rederechtsvergabe in Gesprächen mit dem Benutzer. Insbesondere überlappende Sprache und Unterbrechungen sind mit Vorurteilen über Dominanz, aber auch mit positiven Phänomenen wie gemeinsamer Begeisterung verbunden. Stille kann peinlich sein oder ein Zeichen von geduldigem Zuhören.

Um konsistentes Verhalten für ein breites Spektrum von Gesprächsagenten - vom gehorsamen Haushaltsgehilfen bis zum Antagonisten in Trainingssimulationen - zu erzeugen, ist ein psychologisch fundiertes Modell erforderlich. In dieser Dissertation wurden daher bestehende Theorien darüber herangezogen, wie sich Persönlichkeit und zwischenmenschliche Einstellung in der Rederechtsvergabe widerspiegeln, insbesondere im Timing der Sprechaktivität und den begleitenden Blicksignalen. Anschließend wurde ein entscheidungstheoretischer Ansatz gewählt, um idealisierte, menschenähnliche Denkprozesse zu modellieren und so ein Gleichgewicht zwischen der Generierung eines natürlicheren Agentenverhaltens und der Erfüllung der hohen Erwartungen zu finden, die Menschen an eine rationale Maschine stellen.

Das hier vorgestellte Konzept besteht aus drei Teilen. Das *Einflussdiagramm* wählt das Verhalten zur Rederechtsvergabe, wie z.B. zu sprechen beginnen oder den Blick abwenden, das die Ziele des Agenten am besten erfüllt. Das *Participant Framework* verbindet dieses Verhaltensmodell mit dem Dialogmanager, indem es Kontextinformationen für die Entscheidungen des Einflussdiagramms bereitstellt und besagte Entscheidungen zur Regulierung des Dialogablaufs verwendet. Schließlich verbindet das *RobotEngine Framework* die Dialoganwendung mit verschiedenen virtuellen und robotischen Agenten in einer Weise, die das Verhalten zur Rederechtsvergabe von der Implementierung der Agenten getrennt hält.

Schließlich wurde das entwickelte Verhaltensmodell in zwei verschiedenen Beispielkonfigurationen getestet. Bei einem nicht-interaktiven Prototyp mit einem vereinfachten Verhaltensmodell sprachen zwei virtuelle Charaktere miteinander, die als separate Prozesse implementiert und auf explizite verbale Kommunikation beschränkt waren. Diese Einschränkung sollte das unvollständige Wissen simulieren, das ein Gesprächsagent von einem menschlichen Benutzer erhalten könnte. Eine Online-Studie bestätigte, dass die generierten Verhaltensweisen zu Persönlichkeitsbeurteilungen führten, die mit der Theorie und verwandten Arbeiten übereinstimmten. Anschließend wurde ein interaktiver Prototyp mit inkrementeller Spracherkennung und Blickerkennung entwickelt. Eine erste Evaluation wurde durch die Analyse von Aufnahmen von Mensch-Agent-Gesprächen durchgeführt. Während die generierten Verhaltensvariationen mit den erwarteten Mustern übereinstimmten, brachte die Interaktivität zahlreiche Herausforderungen mit sich, die in Zukunft eine gründlichere Analyse erfordern werden. Dennoch wurden wichtige Erkenntnisse gewonnen und als Empfehlungen für die Evaluation eines solchen Systems zusammengefasst.

Acknowledgments

It has been a long journey getting to this point. Over the years spent on this thesis, I have met a lot of people who contributed to it one way or another – with advice, practical help, or much-needed moral support.

First of all, I want to thank my supervisor, Elisabeth André. For not giving up on me when the depression had me in its grip. For providing me with the opportunity to work with a wide range of different robots, and for letting me put my name out there via international conferences, book chapters, or various events for the general public. For encouraging me to drag the plush of my virtual pet around as a memorable conversation starter. For fostering an atmosphere that created close professional and social bonds between staff members.

Next, I want to thank the present and former colleagues whom I met at the chair of Human-Centered Multimedia, as it was called when I began working here. Many of them helped with countless administrative or technical issues, recommended literature that they had come across, gave career advice, or simply asked questions that forced me to put my thoughts into coherent words. Still, there are some whom I want to mention explicitly.

- Markus Häring, who introduced me to the NAOs and the RoboKind R-50s. In particular, the recommendation to keep the different robots interchangeable saved me a lot of trouble.

- Gregor Mehlmann, from whom I "inherited" the Visual SceneMaker setup. We managed to create some cool stuff together.

- Birgit Lugrin, who provided a lot of advice regarding academia, shared the love for the R-50 robots and gave me access to an unexpectedly large pool of study participants for evaluating my first thesis prototype. I am also very grateful for the opportunity to contribute to "The Handbook on Socially Interactive Agents".

- Andreas Seiderer, who was often the one to pull me out of my office in the evening, or vice versa. Thank you for the long conversations about all sorts of topics – including computing hardware, technical advice, the academic system in general, 3D printing, or our favorite video games.

- Hannes Ritschel, my fellow robot wrangler. For sharing the responsibilities for those pesky little things that tended to fail at the worst possible moment, and the cool moments when they actually worked. And for sharing a lot of creative interests, both outside the lab and in those areas where work and hobbies overlapped.

- Simon Flutura, another highly creative spirit. For showing me the importance of not taking things too seriously and finding a proper balance between the technical and artistic parts of my mind.

- Katharina Weitz, for providing psychological expertise and help with designing studies for various projects during that time. Also, for in-depth discussions about academic life or career strategies, and for lifting my spirits over lunch or a cup of coffee.

- Pooja Prajod, the colleague sharing my office during the last years. For countless conversations about our respective research topics and our academic careers in general – and for making sure both our brains were sufficiently fueled with coffee before tackling the second half of the workday.

I also want to thank various researchers who visited our lab over the years, took the time to listen to my plans and provided feedback or helpful pointers. These include Charles Rich, Candace Sidner, Brigitte Krenn, Leo Wanner, Cristina Conati, Sharon Oviatt, and Patrick Gebhard. Furthermore, I met countless people at the conferences, workshops and project meetings I attended - people who provided scientific advice, became valuable contacts, or provided much-needed emotional support during my first public presentation of my thesis topic. As for the people whom I only ever met online, I specifically want to thank Thilo Michael for getting me started on using the Retico framework for my interactive prototype.

The next round of thanks goes to my social support system outside of work.

- My father and mother, Rainer and Silvia Janowski, for supporting my career choices ever since kindergarten when I first wanted a toy robot.

- My sister Michaela for sharing the burden of caring for said parents, for long conversations about the ups and downs of university life, and for sometimes just getting silly in between all that. Like when I returned

from a stressful business trip to find one of my plushies wearing a hand-knit sweater.

- My aunt Irene and cousin Dominik, especially for inviting me to the DrachenFest years ago. LARPing turned out to be incredibly helpful for detoxing my brain during vacation.

- Alexander Hollinger, my sword trainer for most of those years, and various fellow students from that school. I am grateful for all that I learned during that time. Swordfighting provided exactly the combination of physical activity and meditative introspection that I needed, and I often found myself noticing parallels between my thesis topic and the coordination between training partners. Furthermore, the social gatherings were welcome breaks between periods of work.

- Sabine Leimer, my recorder teacher since I was 7 years old. We have long since moved on from a teacher-student relationship, and I am very grateful that our group still finds semi-regular appointments for playing together, despite all the ways in which life changed since back then.

- Joschi Hofmann, my guitar teacher for several years. Thank you for patiently listening to my ramblings about defective robots, publications in the making, and whatever else university life was throwing at me. It's a pity that I eventually had to quit those classes due to a lack of time, but I am looking forward to picking up the guitar again after this thesis is submitted.

- Diana Stöckert alias Alka-Di-Kijarr, my artist friend for the last couple of years. Thank you for constantly cheering me on (or sternly sending me back to work, when necessary), supporting my mental health recovery, and inspiring me in more ways than I can list here. I am looking forward to where all those ideas will lead us, now that this massive project is finally done.

- Several members of the community that Alka gathered around her INZENTIA project, especially The First Squad. Thank you all for keeping me company on stressful days, listening to my rantings, and celebrating my progress with me.

Unfortunately, in the middle of the thesis, my mental health took a turn for the worse. I was diagnosed with depression and had to be hospitalized for occupational burnout. Therefore, I finally want to thank all the counselors, therapists, and medical professionals involved in putting me back together.

As with the mountain hikes during the lab retreats, the journey towards this point was often steep and rocky, but the view from here is great.

Statement and Declaration of Consent

Statement

Hereby I confirm that this thesis is my own work and that I have documented all sources used.

Kathrin Janowski

Augsburg, January 2024

Declaration of Consent

Herewith I agree that my thesis will be made available through the library of the Computer Science Department.

Kathrin Janowski

Augsburg, January 2024

Contents

Part I

Background

Chapter 1

Introduction

1.1 Motivation

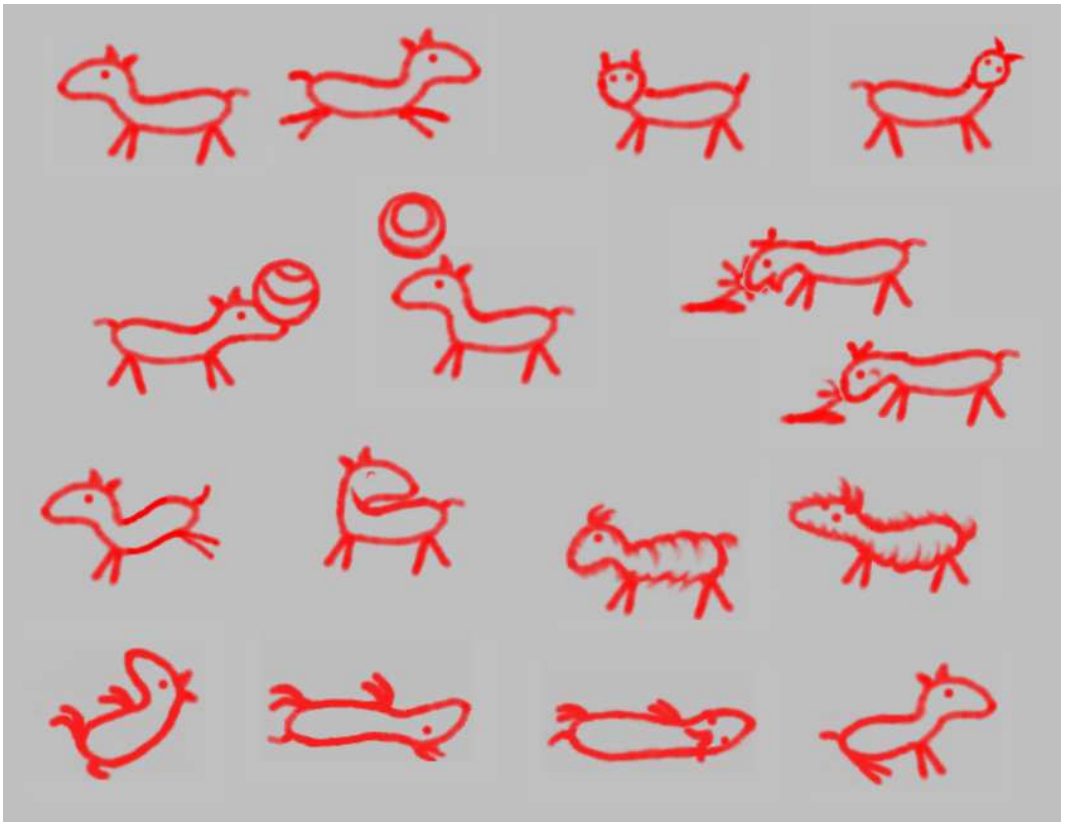


Figure 1.1: Reeya, the virtual pet that I developed during my school years.

Humans have always been fascinated by the idea of bringing their creations to life. In ancient times, there was already the idea of the golem, a servant

made of clay. The often-retold story of Pygmalion sees a marble statue come to life and its sculptor fall in love with it. In the early 19th century, the protagonist of E.T.A. Hoffman's "Der Sandmann" becomes infatuated with a clockwork automaton named Olimpia that its creator presents as his daughter.

With the arrival of computers, powerful new tools became available for this endeavor. Artificial intelligence empowers virtual and robotic creatures to act and react on their own, take on a semblance of life and interact with humans in increasingly natural ways. They speak the user's language, smile or frown, nod or shake their head, and make eye contact or look at things with what seems like curiosity.

More and more computer-controlled characters are becoming part of our daily lives. They serve as actors in entertainment, as non-player characters in video games and animatronic puppets in museums and theme parks. Virtual coaches teach us about healthy lifestyles or help us train for job interviews. Disembodied voice assistants like Siri, Alexa, and Cortana offer to manage our daily life, reminding us of appointments, researching information for us, or playing music at our command. Finally, numerous start-up companies and crowdfunding projects keep promising to bring social robots like Jibo, Buddy, or Olly into our homes that are commonly advertised as potential family members brimming with personality.

1.1.1 Artificial Social Competence

What all these agents need is the ability to communicate with humans in an intuitive, fluent manner. They must know when they are allowed or expected to talk, when they have to yield the conversational floor, or when they need to stand their ground to ensure that an important message is delivered.

Much of this turn-taking is socially coded, whether the designer intended it or not. Humans have evolved to communicate with and relate to other people since prehistorical times. Besides speech, there are numerous other modalities that are used to express opinions and attitudes, show interest, and coordinate when all of these messages can be exchanged.

As a consequence, we tend to see social signals everywhere, not just in other people's behavior but also in that of animals or even inanimate objects. In 1944, Heider and Simmel [52] showed that even the movements of abstract geometric shapes are interpreted as intentional actions, emotional expressions, and cues towards the shapes' personality. Reeves and Nass [107] examined users' behavior towards computers in 1996 and found that people gave rather favorable answers when asked to rate a software program on the same machine on which they had tested it before. However, when the rating was done on a different machine or on paper, their evaluation was less positive and more

varied overall. From this, Reeves and Nass concluded that the study participants subconsciously applied the familiar social rules when interacting with a computer and therefore avoided telling their true opinion to its "face."

Unfortunately, the fact that humans treat artificial entities as social characters also implies that they expect those to follow human behavior rules. A violation of said rules can therefore come across as an intentional lack of respect rather than a technical limitation, even more so when the target audience is unfamiliar with the technology behind the interface. To avoid alienating or offending the users, developers of interactive computer systems need to be aware of the personality that its behavior implies and to shape the user's expectations by making it intentionally display related cues.

Furthermore, not every scenario calls for a compliant personality. A coaching agent that is supposed to assist with behavior change may need to display a certain level of authority, and some users might respond better to a stern "drill sergeant" than to a lenient advisor. Another important application category is that of training simulations. Nowadays, embodied conversational agents serve as role-playing partners that prepare users for job applications [102], negotiation with locals [38, 135], or medical interviews with patients [96]. Here, different personalities allow for increasing or decreasing the difficulty for the trainee. An agent that dominates the conversational floor and refuses to hear the user out presents a greater challenge than one that politely backs down when the user starts speaking.

Finally, which turn-taking patterns are appropriate is heavily dependent on context. What makes the agent appear uninterested in one situation may make it appear patient and respectful in another. For instance, people sometimes struggle with finding the right word or sorting their thoughts into coherent sentences. The resulting pauses within a turn were found to be longer than those occurring when a different speaker takes over [121]. For people affected by cognitive impairments, for example, due to Alzheimer's disease, a listener who waits longer than the social norm requires can make the difference between a frustrating experience and a conversation at eye level [116]. Now that smart speakers and social robots are heralded as the solution to demographic change and the lack of care personnel, this form of social competence becomes a crucial core requirement.

1.1.2 Making Appropriate Choices

Turn-taking behavior, at its core, is the result of repeatedly deciding how to react to the interaction partner and how to resolve conflicts that arise from incompatible timing. Deciding between sitting in silence and taking initiative, interrupting and waiting for the other to finish, or continuing one's sentence

and letting the other take over. Sometimes, what appears to be a conflict on the surface actually signifies a successful interaction, for example when two people are so engrossed in a topic that they enthusiastically finish each other's sentences or blurt out the same idea at the same time. Consequently, the agent must decide when to adhere to the rules and when to break them.

These choices could be implemented with statistical approaches trained on human communication examples, those that are nowadays labeled as "artificial intelligence". However, many of those approaches present themselves as intransparent "black boxes" that provide little justification for their decisions. At the same time, there are decades of research on human personality and body language, as well as artistic conventions for expressing personality via those cues. Statistical approaches rarely take those into account, and risk latching onto irrelevant features that happened to co-occur with the desired choice in the training data. For example, Lapuschkin et al. pointed out examples of image classifiers relying on a watermark or on uniformly-colored padding areas to infer the photograph's subject [74].

The solution suggested in this thesis is to model the decision process of a human interlocutor explicitly. While human reasoning is rarely objective and logical, decision theory offers a structured approach that takes human goals and preferences into account. Tools such as Bayesian networks and influence diagrams have been established for modeling complex, non-deterministic situations [94, p. 47], drawing conclusions about non-observable context factors [94, p. 3] and predicting the consequences of taking certain actions [94, p. 233].

Put simply, a decision-theoretic approach represents the human thought pattern of estimating risks and weighing the benefit of success against the cost of failure. Such considerations also play a role in turn-taking. For example, someone might balance the gravity of being disrespectful to the speaker against that of delaying an urgent message, or the satisfaction of stating their own opinion against the information gained by listening for longer.

By modeling the interlocutor's goals, the personality from which they emerge, and the actions taken to achieve them, the behavior generation is supposed to become transparent and extendable. The graphical representation is expected to aid with configuring different personalities and troubleshooting unexpected behavior patterns without the need for technical expertise.

1.2 Research Questions

Communicative behaviors are usually linked to the personality that observers ascribe to said individual. As will be explained later, what these observers call "personality" equals the person's (perceived) tendency to act in a particular

manner, such as treating others with respect or remaining calm under pressure.

One overarching question arises for modeling turn-taking mechanisms in artificial characters: ***How do an agent's traits influence the coordination of conversational roles?*** This section will split the vague question into three tangible goals at different levels.

1.2.1 Theory

Many psychological models describe similar concepts but still offer different explanations for human behaviors. Much effort has gone into researching the connection between personality and verbal or non-verbal behaviors, while turn-taking has more often been linked to interpersonal attitudes like dominance. Some works describe turn-taking behaviors out of context whereas others link gaze behavior directly to personality or attitudes without controlling for differences in speaking patterns. The relationships between these models need to be identified and put into context to properly aggregate the existing knowledge.

Furthermore, different approaches exist for modeling human or human-like motivations and goal arbitration. Psychological literature focuses on abstract long-term goals [29, 129] whereas research on computer-controlled agents has a stronger focus on concrete actions goals [34], but rarely ties them to an explicit personality. Those that are linked to personality typically use subjective measures for confirming the agent's traits or examining interaction qualities such as "engagement", "likeability of the agent" or "persuasiveness" [89, 3]. Consequently, a link needs to be found between those models, the underlying personality, and - ideally - objective ways to determine goal achievement.

Overall, the driving question for the theoretical part was: ***How does personality translate to variations in turn-taking behavior?*** This question can be split into three steps.

- **How do relevant psychological models relate to each other?** To identify common behavior patterns, it was necessary to unify the existing findings. Concepts such as personality, emotions, relationships, or politeness are intuitively linked, so these connections were expected to help with deriving appropriate behaviors from any of them.

- **How does personality relate to domain-independent goals?** One assumption for this thesis was that certain goals are tied to the functional exchange of messages, while certain social goals are seen across cultures and contexts. Since behaviors can be ambiguous on their own, the underlying intention was seen as a major pillar of the presented behavior model.

- **Which behaviors are used to achieve communicative goals?** Said ambiguity in turn-taking behaviors was approached by associating them with concrete goals rather than the personality itself. This connection is necessary for reasoning about the consequences of the agent's actions in a structured manner.

1.2.2 Concept

A decision-theoretic approach was chosen for the agent's reasoning because it represents an idealized version of human decision-making. However, several questions needed to be answered for translating human mind states and thought processes into a computational model.

For example, there has been research on inferring a user's turn-taking intention from their non-verbal behavior [17], or on predicting how sensitive they would be to interruptions in the current context [56]. These works provide a solid foundation for the presented computational model, but they did not take personality into account. They also rely on subjective measures for the effect of sub-optimal agent behaviors, which is common practice but not recommended for structured decision-making [1]. In a similar way, trade-offs between different costs and benefits are hinted at by literature on personality, but there are no concrete numbers for translating such preferences to weight factors.

The overarching question for the conceptual part was: ***How can the connections between personality and behavior be modeled with a decision-theoretic approach?*** It encompasses the following subtopics.

- **Which uncertainties need to be considered?** Human communication is rarely deterministic. Even if the user were to adhere to a strict script, there would still be issues such as reaction times or the aforementioned ambiguity in producing and interpreting nonverbal behaviors. Predictions of future states are not always reliable, and context variables can only be included up to a certain level of detail before the model becomes too complex. Therefore, the most relevant sources of uncertainty needed to be identified, and appropriate probability distributions needed to be found.

- **What is the utility of communicative behaviors?** Ideally, reasoning about behaviors is grounded in objectively measurable success or failure. However, communicative goals tend to be very subjective, especially when it comes to associating them with a certain personality. Therefore, abstract goals had to be decomposed into concrete sub-goals which the agent's actions could fulfil.

- **How are the utilities traded off against each other?** Literature suggests that humans have the same basic goals, but place different emphasis on them under different circumstances. Those include their culture's social norms, situational constraints such as the performed role, or affective factors such as an individual's personality or temporary emotions. In order to weigh the benefits of an action against its costs, it is necessary to model those influence factors. In this thesis, the focus is on personality.

1.2.3 Implementation

The final step in developing this computational behavior model is to test whether it actually produces the intended behavior. Moreover, the fundamental intention is to use it for improving human-agent communication, so a practical application is the logical conclusion. While several works describe similar uses of decision-theoretic models for selecting agent behavior, they are used to prepare an action sequence [14], plan a waiting time in advance [17, 18] or work on a coarser time scale of minutes or hours [35]. In contrast, incremental dialogue systems rely on statistical methods and time thresholds for choosing the correct timing [38, 30]. Although some reason about the agent's goals, such as wanting to comprehend or to participate [135], the question remains how a decision-theoretic approach can be integrated in a similar setup.

Consequently, the core question for the implementation part was: ***What is necessary for implementing personality-based selection of turn-taking behaviors?***

- **How is the behavior model integrated into the dialogue flow?** The decisions of the model need to be communicated to a dialogue manager so that the agent can take, hold, or yield the turn at the appropriate moment. This must be done as efficiently as possible to support real-time interaction. The model also needs access to enough context information to make informed decisions, but it must be sufficiently independent of the semantics so that it can be reused in different interaction domains.

- **How does the agent keep track of context information?** Although the behavior is decoupled from domain-dependent semantics, there are still many contextual variables that need to be monitored continuously. For example, the agent needs to know what the interaction partner is doing or how far it has progressed in uttering its own contribution. These observations need to be shared with the behavior model at appropriate moments while avoiding computational overhead that would cause

additional delays. Ideally, this information should be stored in a human-readable form that can easily be debugged.

- **How can the behavior be executed on different agent platforms?** Virtual and robotic agents vary significantly in their capabilities and the degree to which manufacturers expose these capabilities to application developers. To keep the behavior model reusable, behaviors need to be defined on an appropriate level of abstraction that nevertheless constrains their observable appearance on the target platform.

- **What is needed to use this behavior model in an interactive setup?** Additional challenges arise every time a computer-controlled character needs to interact with a human in real time. While two conversational agents running in the same application could theoretically share the same "mind" and thus coordinate their turns perfectly, this is impossible when talking to a user. Furthermore, real-time input recognition typically demands a lot of computational resources on its own, so the inference of the optimal turn-taking behavior needs to be as efficient as possible. Finally, evaluating such a setup is not straightforward because human behavior is hard to control and reproduce in an experiment.

1.3 Outline of this Thesis

This thesis offers a new approach to modeling turn-taking for embodied conversational agents. While answering the questions above, it aims to provide a transparent, configurable, and re-usable behavior model that can combine hand-crafted rules with machine-learned parameters. The document is organized as follows.

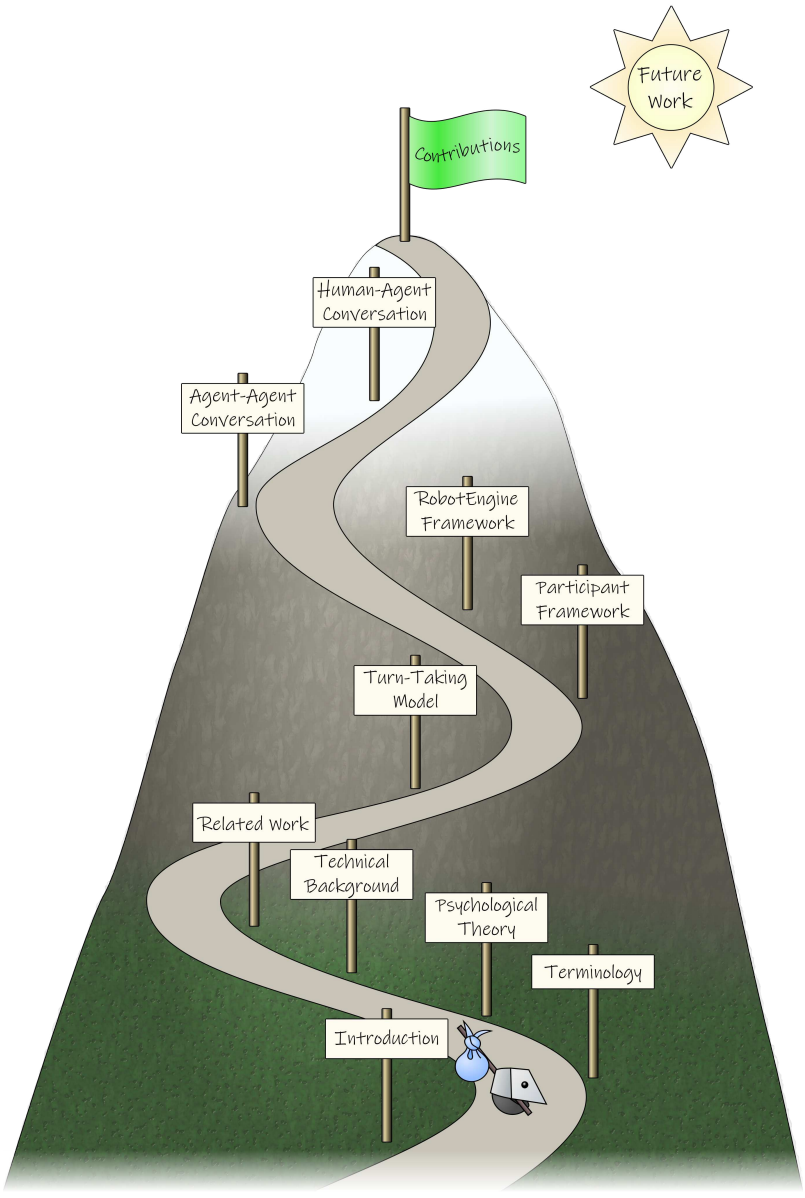
The first part will lay out the theoretical and technical background. Chapter 2 will define the relevant terminology and compare how certain concepts are defined in psychology and computer science. It will be followed by a chapter on the psychological models informing this research, the goals and motivations behind human interaction, and the means through which they coordinate their communicative efforts. The next chapter will focus on the technical background, covering topics such as the fundamentals of decision theory or existing agent frameworks. Afterward, related work in computer science will be presented. There, the focus will be on approaches for action timing, the mapping between mental states and observable behavior, as well as adaptable and adaptive human-agent interaction.

The second part of the thesis presents the chosen approach. It contains chapters about developing the turn-taking model, extending the chosen dia-

logue manager to use it, and realizing the selected behaviors on different agent platforms.

The third part focuses on proof-of-concept applications built around this turn-taking model. One chapter describes a simplified setup, using two independent agents that act out different personality profiles in front of a passive observer. Another chapter is dedicated to an interactive human-agent setup and the additional challenges of evaluating a real-time dialogue application.

The thesis will conclude with an outlook on the research landscape. After a chapter that summarizes the contributions, the final chapter will discuss the limitations of this research and point out directions for future work.



Chapter 2

Terminology

2.1 Introduction

To transfer human communication patterns to computer-controlled agents, it is necessary to combine findings from psychology with established practices from human-agent interaction. However, this combination of different scientific disciplines is challenging.

Social sciences use different terminology than computer science. Sometimes, they use different terms for the same concept. On other occasions, they use seemingly identical terms, but for different concepts. Therefore, to avoid confusion, the terminology used in this thesis will be clarified before everything else.

This chapter contains sections on four major topics. Conversational agents are the main focus of this thesis, so the related concepts are defined first. Next, there will be definitions related to their goals and intentions because the presented approach revolves around what the agents want to achieve by choosing one behavior over another. The third section focuses on communication, the agents' primary task. Finally, several affective concepts will be defined because this thesis aims to model not only context-appropriate behaviors but especially behaviors that portray specific agent personalities.

Each of these sections summarizes important definitions from the different scientific disciplines before defining the terms that were chosen for the research presented here.

2.2 Agents

According to the Merriam-Webster dictionary[1], the word "agent" is derived from the latin word "agere" that has meanings along the lines of "to drive cattle", "to be in motion", or "to perform". In the broadest sense, an *agent* is an entity that acts.

2.2.1 Definitions in Computer Science

Autonomous Agent

An *autonomous agent* is a piece of software that acts on its own, usually to assist a human user by carrying out direct orders or acting according to their (assumed) preferences. As Cohen and Levesque [33] put it, such an agent has a "mental state" or "web of beliefs" that can be influenced by its own actions and those of others. Furthermore, an agent intends to achieve something in accordance to those beliefs, for example, to change the world state [105] or the beliefs of a different agent [34]. More information about these beliefs and world states will be provided in section 4.3.1.

Embodied Conversational Agent

An *embodied conversational agent* is designed to interact with a human via verbal and nonverbal behaviors. This term is mostly used for graphically displayed agents. For example, the GRETA platform [95] was developed to control an animated 3D character that can communicate via synthesized speech, facial expressions, hand gestures, or gaze direction. Some of those agents are limited to an expressive head, such as the quizmaster agent implemented by Bohus and Horvitz [18], whereas others are designed to mimic humans as closely as possible [51].

Social Robot

In contrast to industrial robot arms or tools like robotic vacuum cleaners, a *social robot* is designed to mimic a living being and communicate with people. They can be built to resemble humans, animals, or fantasy creatures with varying degrees of stylization. Such robots are typically used in similar applications as graphically embodied conversational agents.

For example, Leite et al. used the Philips iCat robot, which has a vaguely cat-shaped body with a cartoon-like face, as an emotional game-playing companion for children [77]. The Aldebaran NAO, which has an articulated hu-

[1] https://www.merriam-webster.com/dictionary/agent#word-history

manoid body but no expressive face, is well-established in the global research landscape [76, 83, 4]. The Furhat robot used by Skantze et al. [122] combines the advantages of a robot's physical presence with the flexibility of a graphically animated face on a mechanically articulated neck.

2.2.2 Definitions in this Thesis

Embodied Conversational Agent

This thesis will use the term embodied conversational agent (ECA) for both graphically displayed agents and social robots. The argument here is that both are used in dialogue-based interfaces and that very similar principles apply for transferring human communication behaviors to them.

Computer-controlled Character

The term "computer-controlled character" will sometimes be used interchangeably with "embodied conversational agent". Besides avoiding repetitive text, this term also emphasizes that the agent portrays a character with a specific personality, like an actor. The words "computer-controlled" also stress that the agent is autonomous as opposed to remote-controlled by a human. Furthermore, this term was chosen for communication with the general public who are less familiar with the terms "agent" or "embodied".

2.3 Goals and Intentions

The psychological literature on human motivations tends to focus on long-term, high-level goals, whereas goals for computer-controlled agents are more often defined in a pragmatic, short-term way.

2.3.1 Definitions in Psychology

Motivation

Barbuto and Scholl [15] define *motivation* as rooted in various *desires*. They distinguish between different motivation sources depending on what exactly is desired. Talevich et al. [129] defined *motives* as "things that people want" and that consequently drive them towards specific behaviors.

Goal

Chulef, Read, and Walsh [29] refer to *goals* as "motivational entities" and appear to use the term interchangeably with *motivation*. However, Talevich et

al. [129] made a point of distinguishing between *motives* and *goals*. According to them, the latter represents an "end state" that can either be the final result of one's actions or an intermediate step toward said result.

2.3.2 Definitions in Computer Science

Goal

Cohen and Levesque [33] define an agent's *goal* as one of its *desires* that it has actively chosen to pursue. More specifically, such a goal is a *world state* that the agent prefers among all possibilities. They further distinguish between *achievement goals* and *maintenance goals*, depending on whether they are abandoned after the associated world state becomes true. Finally, *persistent goals* are defined as goals that are pursued until the agent believes it to be achieved or impossible, or until an *escape condition* is met that allows the agent to abandon that goal.

Intention

Cohen and Levesque [33] define *intentions* as a special type of persistent goal that is only achieved if the agent believes that it performed an action that causes the desired world state. Additionally, this action needs to be preceded by the agent's belief that it will have done said action in the future.

2.3.3 Definitions in this Thesis

Goal

In this thesis, the distinction between goals and intentions is less clear-cut than in related works. Partially this is a simplification, and partially this is because the line between the agent's desires and the actions taken as a consequence is drawn differently.

For instance, the final version of the developed behavior model (see sections 6.2.3 and 10.2.1) contains a goal named "speak" which is separated from the agent's decision to produce audio for transmitting the current sentence. More precise names for that goal would be "be in a speaking state", "have the floor", or "be using the audio channel". Likewise, the goals "hear" and "see" would more aptly be named "be able to hear" or "be able to obtain visual information". However, such names would be very cumbersome in both the influence diagram's depiction and the text describing it.

The agent's actions, on the other hand, are defined on a very low level. While they can be named along the lines of "speak" and "wait" (as they are in the simplified non-interactive prototype in chapter 9), they more accurately

correspond to permitting and prohibiting the use of the associated modality. The intention to actually speak is formed and executed outside the behavior model, as will be explained in sections 7.4.3, 9.3.2, and 10.3.1.

Therefore, the term *goal* in this thesis refers to something that the agent wants to be doing after the point of decision.

Communicative Goal

A *communicative goal* in this thesis is not defined in terms of the communicated content, but rather the ability to communicate said content. In contrast to the definitions of Cohen and Levesque [34], this thesis does not focus on altering the interlocutor's belief. While the behavior variations generated by the suggested turn-taking model are meant to evoke a certain belief regarding the agent's personality, the agent itself does not directly aim to appear introverted or agreeable. Instead, the agent's goals relate to the process of sending and receiving messages. The personality perception is treated as a side effect that emerges from the agent's priorities.

Goal Conflict

This thesis will consider it a *goal conflict* when one participant has to weigh one of its goals against one or more others. While one could argue that a *turn-taking conflict* (see section 2.4.3 represents a conflict between both participants' goals, one participant's goals are not directly accessible for the other. Consequently, a participant can only reason about their own goals, which may or may not include the goal "help the interaction partner achieve what I assume is their goal".

2.4 Communication

Computational modeling of communicative processes often uses the established terminology from the social sciences. There is a notable intersection between both disciplines because research in human communication often relies on software tools for systematically analyzing observations while computer scientists seek to transfer the results to ECAs.

2.4.1 Definitions in Psychology

Floor

The person who is currently speaking is said to have or hold the *floor* [70, 41, 113]. When they stop speaking, they yield said floor [41] and when they start,

they take it [70, 41]. Interlocutors can request the floor [41], offer it to another person [70], or otherwise negotiate for it.

Turn

During a conversation, people tend to take *turns* [41]. They usually speak one after the other, although *simultaneous turns* can happen accidentally or be permitted under certain circumstances. Duncan [41] also distinguishes between *turn* and *speech* because listeners can speak without the intention to take the floor from the other person.

Back Channel

Listeners tend to give a variety of short reactions during the speaker's turn. These so-called *back channels* include nodding, acknowledging comments like "okay" or "hmm-hmm", or completions of the interlocutor's sentence [41][71, p. 260]. According to Duncan [41], those do not count as a turn on their own but rather signify that the listener avoids taking the conversational floor. In other words, they encourage the speaker to continue [71, p. 260].

Interruption

Clark [31] explains that people usually follow the rule of speaking one after the other, switching either when the floor is handed over or when the first speaker has finished speaking. According to him, *interruptions* are intentional violations of that rule. They happen when a different person starts speaking during the current speaker's turn and the latter decides to stop speaking in response to this.

According to Goldberg [48], interruptions are typically seen as "an act of conflict, competition, or non-involvement". More details on the interpretation of overlapping speech will be provided in section 3.4.3.

2.4.2 Definitions in the Intersection between both Fields

Research in human communication requires the careful analysis of observations. Therefore, annotation standards are often developed alongside these research efforts, together with software tools and data formats that facilitate the analysis of recordings.

At its core, an interaction consists of the exchange of messages in different channels. These messages need to be described in a consistent way, not only by researchers studying human communication but also by those developing

software for natural language understanding (NLU) and autonomous conversational agents.

Communicative Act

Bunt et al. proposed the Dialogue Act Markup Language (DiAML) [23, 22] for annotating human conversations. It was formalized as the ISO 24617-2 standard [60] in 2010 and revised in 2020. In this standard, they defined a *dialogue act* as the semantic information attached to a *functional segment*, a minimal part of the observed behavior that can transmit meaning.

These functional segments can be verbal or any other combination of modalities [101]. For example, the function "accept" can be communicated by saying "okay", nodding, or doing both. The information that someone is paying attention can be transmitted with the words "I'm listening" or by looking at the speaker.

Communicative Function

Each act in DiAML has a *communicative function* that refers to the effect that a message has on the addressee. For example, they could interpret the sender's words as a request or simply as a statement providing information. This function is not necessarily tied to the linguistic form [22]. For example, certain offers or requests are phrased as questions, such as "Would you care for some tea?" or "Do you know what time it is?"

Content

In addition to the communicative function, many communicative acts have semantic *content*. The content is the information that the addressee is meant to process according to the communicative function. For example, this could be the topic about which the sender requests more information or the action that they want the addressee to perform.

For example, a greeting or an expression of gratitude does not necessarily need content in order to advance the conversation. Nevertheless, these acts could come with useful information such as the specific thing that someone is grateful for.

2.4.3 Definitions in this Thesis

Communicative Act

This thesis will use the definitions of *communicative act*, *communicative function*, and *content* as described in section 2.4.2.

Message

In this thesis, a *message* will be the verbal or non-verbal signal that carries a communicative act. This can be a spoken sentence, a gaze shift, or the beginning and end of a participant's voice activity.

Turn-Taking Conflict

Following the typical interpretation of interruptions, this thesis will consider it a conflict when two interlocutors try to speak at the same time. This definition will also include unwanted silence that stems from both participants choosing to wait. In other words, a *turn-taking conflict* happens when both sides simultaneously try to make the same turn-taking choice - speaking or waiting.

2.5 Personality and Interpersonal Relationships

As with communication research, computer science typically builds on psychological definitions when simulating personality, emotions, or social relationships in human-agent interaction.

2.5.1 Definitions in Psychology

Personality

According to Argyle and Little [11], *personality* refers to the behavior tendencies that somebody displays rather consistently across time and situations. If they find themselves in a particular combination of role, task and observers, people are likely to react in a way that these observers can learn to predict. This includes systematic variations based on context factors. For example, a colleague at work sees a different side of the same individual's personality than a family member in the privacy of that same individual's home. The colleague and the family member may attribute different personalities to the observed person, but their respective perceptions will largely remain the same.

McCrae and Costa [80] point out that the five dimensions of the widely-used "Five Factor Model" for personality strongly relate to how a person is perceived by others. While only two of those dimensions directly influence one's relationship to others, the remaining three still determine emotional responses to situations and the approach to role-based responsibilities.

A closely related term is *temperament*. Mehrabian [87] defined it as a person's disposition towards experiencing particular emotions or patterns thereof "across representative life situations". As with the personality definition by

Argyle and Little [11], the key aspect is the large number of situations in which that person responds in a similar way.

Details on the "Five Factor Model" and Mehrabian's temperament model will follow in sections 3.2.2 respectively 3.2.1.

Interpersonal Relationships

Besides personality, much research has focused on interpersonal adjectives [138] and behaviors [99]. Some works speak of "interpersonal traits" [80] or "dimensions" [138] that are used to describe "interpersonal dispositions" [80, 142] and "relationships" [125]. Those, in turn, result in the aforementioned behaviors. Details on those interpersonal dispositions and their connection to personality will follow in section 3.2.3.

Politeness

In their "Politeness Theory", Brown and Levinson [20] defined two primary strategies for being polite. They proposed that every human has two fundamental needs regarding their public self-image, called "face" in that work. Those needs, called "positive face wants" (need for appreciation) and "negative face wants" (need for autonomy), can be threatened by somebody else's actions. In order to avoid this, people use "positive politeness" and "negative politeness" that accommodate the respective face need. More details on these needs will follow in section 3.3.1.

2.5.2 Definitions in Computer Science

Personality

Research on the computational simulation of personality builds on the psychological definitions for this concept, in particular the notion of consistency over time. Ball and Breese [14] pointed out that an agent's behavior needs to be consistent because personality changes very slowly or not at all. Gebhard [44] defines personality as "the set of a person's or a virtual character's features that are relatively stable over time and allow for distinguishing between them" (p. 15, translated from German).

Interpersonal Attitude

Works about the behavior of autonomous agents, towards other agents as well as towards humans, speak of "interpersonal attitude" [130, 106, 47]. Said attitude is defined in terms of the same dimensions that psychology uses to describe interpersonal dispositions and relationships.

2.5.3 Definitions in this Thesis

Personality

This thesis uses the established definition of personality as behavior tendencies that stay rather stable over time and across situations. The agent's personality traits are meant to be provided by the interaction designer and stay the same throughout the interaction[2].

Interpersonal Attitude

Related works in computer science tend to use "interpersonal attitude" rather than "interpersonal traits" or "interpersonal relationship". Therefore, the first term will be adopted for this thesis.

Politeness

Politeness will be defined according to Brown and Levinson's politeness theory [20]. Their definitions of *positive/negative face*, *face wants* and *positive/negative politeness* will be used as described in section 2.5.1.

2.6 Conclusion

This chapter summarized the most relevant terminology in social sciences and computer science. Furthermore, it defined the terms that will be used in this thesis, drawing definitions from both disciplines and comparing them to established terms where necessary. The following two chapters will explain most of those concepts in more detail.

[2]There will be a short discussion of adapting the personality to the user in section 12.3.4, suggesting ways to adjust the configuration at runtime. However, this is outside the scope of this thesis.

Chapter 3

Psychological Theory

3.1 Introduction

To ensure that machines and humans "speak the same language", we first need to understand how human communication works.

Humans use several channels for transmitting messages, such as speech, gaze, body pose, or facial expressions. Many of these transmit additional nuances in the way the message is formed.

A smile may be seen as an indication of happiness, which in turn hints at satisfaction with the present situation. Lowering the head may be read as sadness or shame, indicating failure. Directing a prominent pair of eyes towards an object implies the desire to inspect it more closely and may be labeled as "curiosity". Interrupting the speaker is often taken as a sign of impatience or a desire to control the other.

Besides delivering the actual message, people also need to coordinate their communication process with each other [32]. Like all behaviors, the coordination mechanisms contribute to the opinion that an observer forms about the interlocutors. In some cases, for example with gaze, it is also hard to disentangle explicit coordination signals from the underlying attention [8, p. 170], which in turn might be rooted in liking the other person [8, p. 58-63] or be part of a learned politeness pattern [8, p. 29].

Personality refers to the tendency to display certain behaviors or emotional responses under certain circumstances [11, 87]. Such a tendency, in turn, allows an observer to predict future reactions in similar contexts. Relationships form another facet that is intuitively connected to personality. The traits of extraversion and agreeableness are often defined in terms of social interaction, such as the tendency to approach others or to cooperate with them.

This chapter will present the psychological background for the coordination mechanisms used in conversation. First, section 3.2 will give an overview of personality models and the traits that will later become the input parameters for the computational model developed in this thesis. Section 3.3 will then identify specific interaction goals that stem from an interlocutor's personality and interpersonal attitude, and section 3.4 will describe the verbal and nonverbal behaviors that are used to coordinate the actions of the interacting parties. Finally, section 3.5 will summarize and conclude the chapter.

3.2 Personality and Interpersonal Attitude

Psychologists have proposed numerous models for measuring and classifying personality, interpersonal attitudes, and affective states. Often, there is no clear distinction between these three concepts - for instance, the so-called Interpersonal Adjective Scale plays an important role in defining personality traits [134, 53] while other sources define personality in terms of the emotions which people with a given trait are likely to experience [87]. Consequently, there have been approaches to compare and relate some of these models to one another, and to convert representations between them [80, 117, 86].

This section will present several models that have been linked in such a manner and will provide the foundation for constructing a computational model in later chapters.

3.2.1 The Pleasure-Arousal-Dominance Model

One fundamental model for affective states is based on two primary dimensions: The *Evaluation* as pleasant or unpleasant, also known as *Valence* [141], and the degree of *Activation* or *Alertness*, also known as *Arousal*. In 1980, Russell [115] mapped 28 emotion adjectives to a circumplex defined by these two dimensions. His results can be seen in figure 3.1. He argued that affective states were better represented by systematic combinations of these dimensions than by independent emotion categories.

In that same source [115], Russel suggested that there could be a third dimension involved, such as *Dominance*. Earlier, in 1968, Osgood [99] had already concluded from intercultural studies that there were three stable factors that defined a wide range of emotional and interpersonal concepts. He labeled these as *Evaluation*, *Potency* and *Activity*, and referred to them as the *E-P-A system*.

In 1977 Russell and Mehrabian confirmed that these three dimensions were a minimal set of factors capable of describing large numbers of affective states

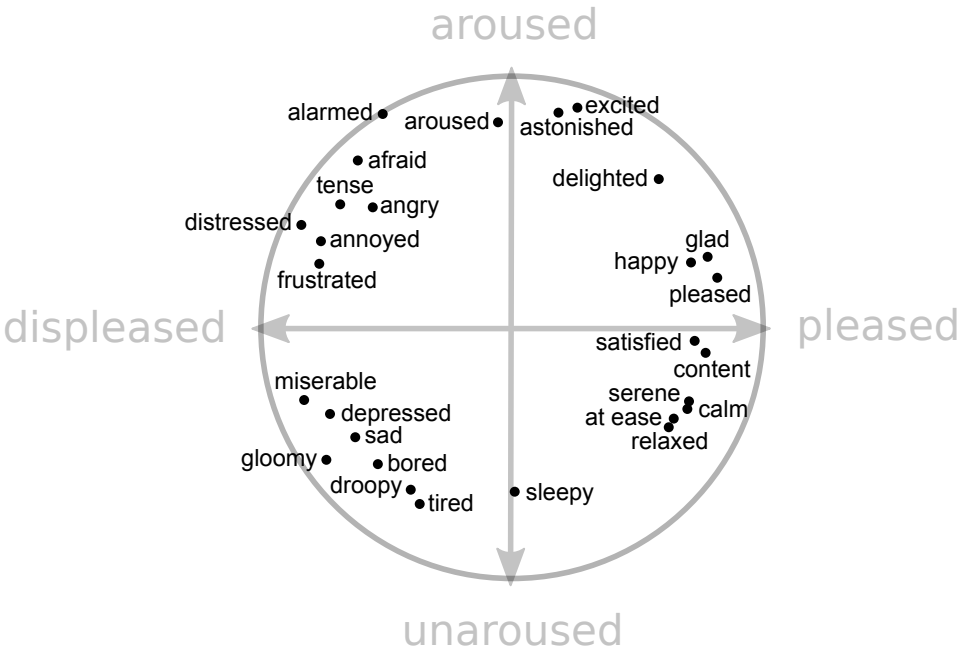


Figure 3.1: The affective circumplex with the placement of adjectives as given by Russell [115].

[114]. In particular, they listed 151 terms with their respective mappings to Pleasure, Arousal, and Dominance. Part of these mappings are shown in figures 3.2 and 3.3.

Mehrabian later consolidated those findings into what he called the *PAD Temperament Model* [87, 86]. He defined *Temperament* as "emotional traits [which] are stable over periods of years or even a lifetime", in contrast to emotional states which "can vary substantially, and even rapidly, over the course of a day". According to him, an individual's temperament corresponds to the average of their experienced emotional states, expressed as PAD vectors.

This definition of temperament resembles the concept of *Personality*. Indeed, Mehrabian reported mathematical relationships between the PAD dimensions and other established personality traits [87, 86], including the Five-Factor Model which will be described in section 3.2.2.

In summary, the three dimensions are defined as follows.

Pleasure

The temperament trait *Pleasure* equals a person's tendency towards positive or negative affect [87].

Osgood [99] defines the synonymous *Evaluation* factor by opposite pairs such as "good-bad", "kind-cruel" and "pleasant-unpleasant". According to Russell and Mehrabian [114], terms such as "thankful", "secure", "modest"

Figure 3.2: Affective terms mapped to the Pleasure-Arousal plane. Coordinates taken from Russel and Mehrabian [114].

and "cooperative" can be found on the positive half of this axis, as well as "friendly", "affectionate" or "in love". Words like "displeased", "regretful", "discouraged", "anguished" or "disdainful" are found on the negative half. Russell's circumplex [115] places "miserable", "depressed" and "frustrated" close to the negative pole whereas "happy", "pleased", "satisfied" and "content" are near its positive counterpart. Mehrabian [87] gave "affectionate-nasty" and "snobbish-generous" as examples of antonym pairs that are equal with regard to the other two dimensions. In the same source, he further reported strong positive correlations with the personality traits of nurturance and agreeableness.

Figure 3.3: Affective terms mapped to the Pleasure-Dominance plane. Coordinates taken from Russel and Mehrabian [114].

Arousability

According to Mehrabian [87], *Arousability* models a person's responsiveness to stimuli and the time it takes for them to return to their normal arousal level.

In Osgoods's E-P-A system [99], the factor *Activity* corresponds to adjective pairs such as "active-passive", "quick-slow" and "excitable-calm".

In PAD space [114], states such as "astonished", "terrified", "enraged" or "excited" are fairly high on the arousal axis, whereas "relaxed", "quiet", "fatigued", "uninterested", "listless" and "bored" represent low arousal. Terms like "tired", "droopy" and "sleepy" can be found near the "unaroused" pole of the Affective Circumplex [115] while "alarmed", "aroused", "astonished" and "excited" are located near the opposite pole. As for measuring this trait,

Mehrabian [87] gives the example phrases "I get happy or sad easily" and "I am not affected much by the positive or negative mood of a crowd". Agreement with the latter is inverted.

Dominance

Mehrabian [87] defines *Dominance* through the degree of control a person believes to have over their situation, as opposed to being controlled by others or external circumstances.

The E-P-A system factor of *Potency* is defined by opposing terms such as "strong-weak", "hard-soft" and "big-little" [99]. Russel and Mehrabian [114] placed adjectives like "bold", "useful", "mighty", "proud" and "powerful" in the positive range of this dimension. The negative range contains words such as "awed", "overwhelmed", "protected", "humble", "shy", "fearful" or "helpless". The statements which Mehrabian [87] used to measure Dominance included "I go my own way instead of following others" and the reverse-coded "sometimes I hesitate to express my ideas".

3.2.2 The Five-Factor Model

A widespread model for classifying personality is the Five-Factor Model, also known as the "Big Five" [134, 81, 86, 49, 104] or the "OCEAN" model. Consequently, their relationship to other models, such as the PAD space (shown in figure 3.4) or the Interpersonal Circumplex (which will be explained in section

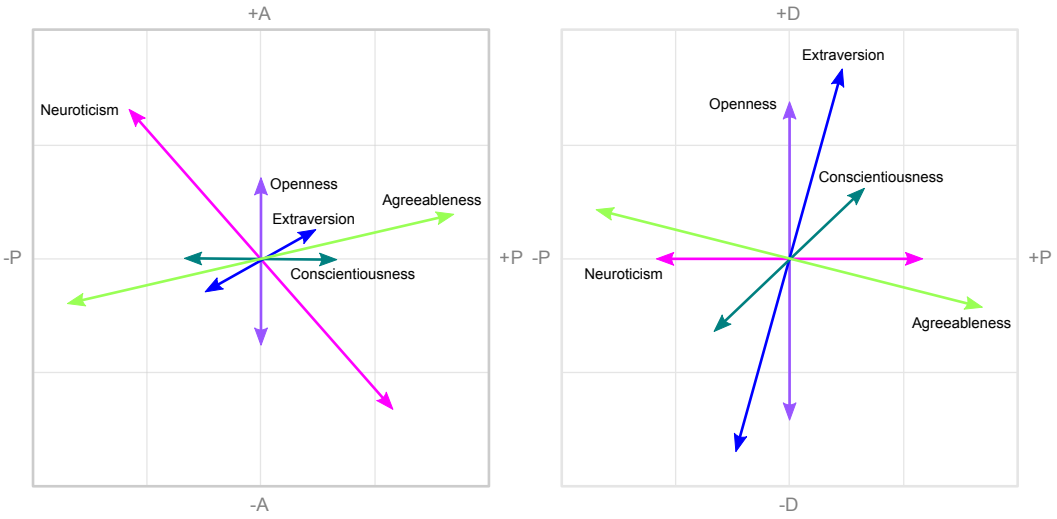
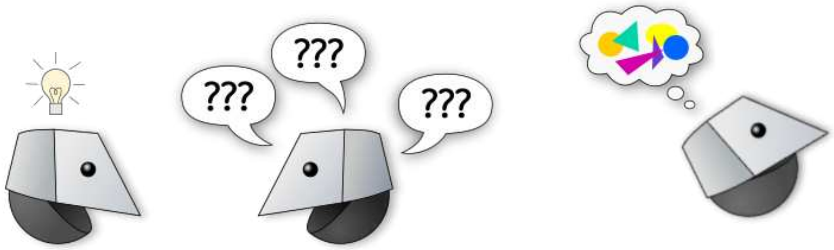


Figure 3.4: Location of the Big Five dimensions in PAD space, according to Mehrabian [86].

3.2.3), has already been examined by psychologists. This makes the Big Five an important foundation for this thesis.

Several questionnaires and item inventories exist for measuring these factors, such as the *Big Five Extension of the Revised Interpersonal Adjective Scale* by Trapnell and Wiggins [134], the *Ten Item Personality Inventory (TIPI)* by Gosling, Rentfrow and Swann [49] and the *Big Five Inventory-10 (BFI-10)* by Rammstedt and John [104].

The five dimensions are defined as follows.



Openness

The *Openness* dimension covers personality aspects associated with being open-minded towards ideas and experiences. It is also known as *Sophistication* or *Culture* [86].

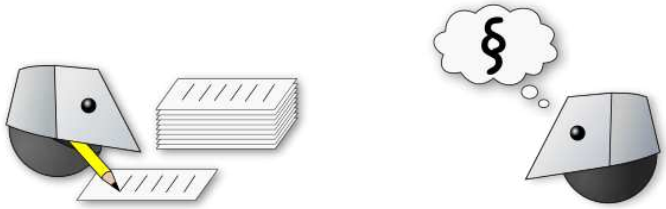
Trapnell and Wiggins [134] identified five main facets of this trait which they labeled as "intellect", "creativity", "curiosity", "reflectiveness" and "unconventionality".

McCrae and John [81] gathered numerous adjectives that are used to describe the trait in literature and questionnaires. According to those, people with high Openness are "artistic" and "aesthetically reactive", "curious", "imaginative" and "original", and think and judge in "unusual" and "unconventional" ways. They also value "intellectual matters", show a "wide range of interests", and are "insightful" and "introspective".

The TIPI [49] measures Openness on two subscales: The item "open to new experiences, complex" and the reverse-scored item "conventional, uncreative". The BFI-10 [104] measures it using the item "has an active imagination" and the reverse-scored item "has few artistic interests".

According to Mehrabian [86], the PAD representation of this trait is:

$$Sophistication = 0.16 Pleasure + 0.24 Arousal + 0.46 Dominance$$



Conscientiousness

Conscientiousness is generally associated with being well-organized, adhering to rules, and fulfilling one's duties.

Trapnell and Wiggins [134] named adjectives such as "organized", "tidy", "neat", "efficient", "thorough", "self-disciplined" and "reliable".

McCrae and John [81] further listed "responsible", "productive" and "dutiful", as well as "competence" and the tendency to "behave ethically". Ambition seems to play a role as well, as shown by the attribution of a "high aspiration level" and the description "achievement striving".

Mehrabian [86] likewise quoted the "will to achieve", but pointed out that this aspect differed from that of being "neat, well organized and diligent". In the TIPI [49], conscientiousness is measured by the items "dependable, self-disciplined" and the reverse-scored item "disorganized, careless".

The BFI-10 [104] uses "does a thorough job" and the reverse-scored "tends to be lazy".

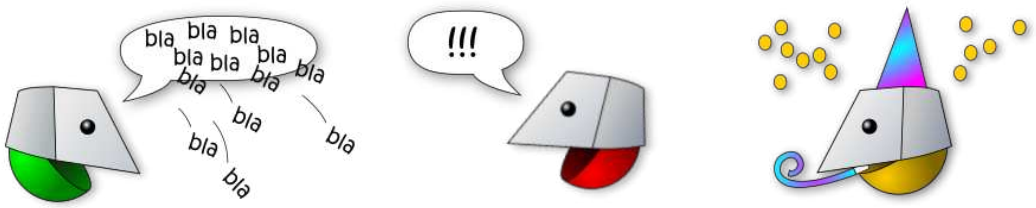
According to Mehrabian [86], the PAD representation of this trait is:

$$Conscientiousness = 0.25 Pleasure + 0.00 Arousal + 0.19 Dominance$$

Extraversion

Extraversion describes a person's tendency toward confident and sociable behavior.

Under the alternative term "Surgency", Trapnell and Wiggins [134] list characteristics like "dominant" and "domineering", "assertive", and "self-confident" for high levels of this factor, and words like "meek", "shy" and "timid"

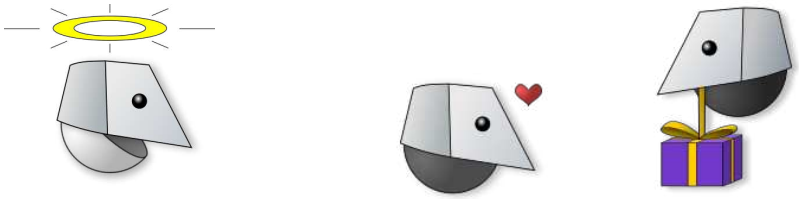


for the opposite pole. McCrae and John [81] add qualities such as "energetic", "talkative", "facially, gesturally expressive", "outgoing", and "gregarious".

Gosling et al. [49] measure Extraversion with the item "extraverted, enthusiastic" and the reverse-scored item "reserved, quiet". Rammstedt and John [104] use "outgoing, sociable" and the reverse-scored "reserved".

According to Mehrabian [86], the PAD representation of this trait is:

$$Extraversion = 0.29 Pleasure + 0.00 Arousal + 0.59 Dominance$$



Agreeableness

The trait *Agreeableness* is associated with being likable and getting along well with others.

Trapnell and Wiggins [134] related adjectives like "kind", "gentle-hearted", "sympathetic", and "accommodating" to high Agreeableness, whereas "ruthless", "uncharitable", "cruel", and "coldhearted" represent low Agreeableness. McCrae and John [81] further list qualities such as being "appreciative", "forgiving", "trusting", and "compassionate", as well as "altruism" and "modesty". The TIPI [49] uses "sympathetic, warm", and the reverse-scored "critical, quarrelsome" to measure this trait, while the BFI-10 [104] uses "is generally trusting" and "tends to find fault with others", respectively.

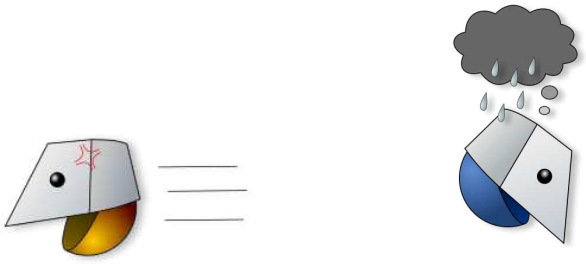
According to Mehrabian [86], the PAD representation of this trait is:

$$Agreeableness = 0.74 Pleasure + 0.13 Arousal - 0.18 Dominance$$

Neuroticism

Neuroticism is the inverse of *Emotional Stability*, which is another common label for this trait. It covers a person's tendency toward negative affect, mood swings, and impulsive behavior.

Trapnell and Wiggins [134] relate high Neuroticism to being "worrying", "anxious" and "nervous", as well as "self-conscious", "high-strung", and "overexcitable". Low Neuroticism is associated with being "stable", "calm", and "relaxed".



McCrae and John [81] further list terms such as "thin-skinned", "brittle ego defenses", "fluctuating moods", as well as "depression", "vulnerability", "impulsiveness", and "hostility".

The TIPI [49] measures Emotional Stability on the subscales "anxious, easily upset" and the opposite labeled "calm, emotionally stable". In the BFI-10 [104], the respective items are "relaxed, handles stress well" in contrast to "gets nervous easily".

According to Mehrabian [86], the PAD representation of this trait is:

$$Stability = 0.43Pleasure - 0.49Arousal + 0.00Dominance$$
$$Neuroticism = -0.43Pleasure + 0.49Arousal + 0.00Dominance$$

3.2.3 The Interpersonal Circumplex

Interpersonal behaviors and attitudes are commonly classified using the so-called *Interpersonal Circumplex* [138, 80, 55, 79, 39]. It is formed by two axes: *Status*, which describes the difference in power between the interacting parties, and *Affiliation*, which describes the degree of social closeness.

The circumplex dimensions are closely related to the Big Five dimensions of Extraversion and Agreeableness. McCrae and Costa [80], Markey and Markey [79] as well as DeYoung et al. [39] confirmed that those two factors can be mapped to the same plane, providing an alternative pair of axes which is rotated approximately 30° to 45° relative to Status and Affiliation. Both pairs of axes are shown in figure 3.5.

As with the Big Five, there is a general consensus on the meaning of the two dimensions, but the terminology differs between sources. In addition, Wiggins, Trapnell, and Phillips divided the circumplex further into octants, of which four correspond to the main axes and four to blends between them. Their *Revised Interpersonal Adjective Scale (IAS-R)* [138] consequently provides eight subscales for measuring them. Another questionnaire is the *International Personality Item Pool - Interpersonal Circumplex (IPIP-IPC)* by Markey and Markey [79], which consists of four descriptive phrases per octant.

The circumplex dimensions are defined as follows.

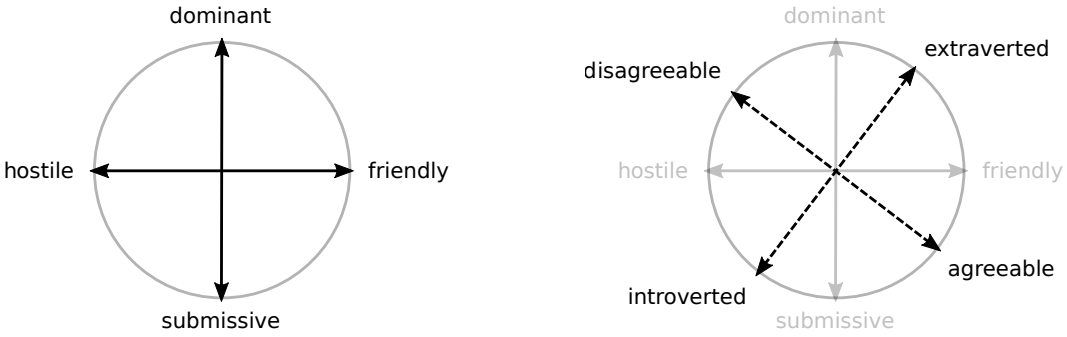


Figure 3.5: The two pairs of axes which define the Interpersonal Circumplex. *Solid:* Status and Affiliation. *Dashed:* Extraversion and Agreeableness.

Status

Status is also known as *Dominance* or *Agency*. This axis passes through the two opposite octants *Unassured-Submissive* and *Assured-Dominant* [138]. It is commonly depicted as the vertical dimension, ranging from *submissive* to *dominant* [81, 55].

The IAS-R [138] measures the positive octant with adjectives such as "self-assured" and "self-confident", "assertive", "dominant" and "domineering". The negative octant is associated with terms such as "timid", "shy", "meek", "unauthoritative" and "unagressive".

In the IPIP-IPC [79], high-status persons are characterized by "demand[ing] to be the center of interest", "demand[ing] attention", "do[ing] most of the talking" and "talk[ing] loudly". In contrast, those with low status "speak softly", "let others finish what they are saying", "seldom toot [their] own horn" and "dislike being the center of attention".

Affiliation

Affiliation is also known as *Love* or *Communion*. Compared to Status, there is less consensus on the labels for this axis. It can range from "hostile" to "friendly" [80], from "indifferent" to "friendly" [55], or from "cold(-hearted)"

to "warm" [134]. The axis passes through the octants "Cold-Hearted" and "Warm-Agreeable" and is commonly depicted as the horizontal dimension.

According to the IAS-R [138], the negative pole is defined by adjectives such as "ruthless", "uncharitable", "cruel" and "unsympathetic", while the positive pole is associated with being "accommodating", "gentlehearted", "sympathetic" and "kind".

The IPIP-IPC [79] relates the "Warm-Agreeable" octant to the tendency to "[be] interested in people", "reassure others", "inquire about others' well-being" and "get along well with others". The subscale for the opposite octant holds the items "believe people should fend for themselves", "don't fall for sob stories", "don't put a lot of thought into things" and "am not interested in other people's problems".

The latter subscale is in line with Horowitz et al. [55] who argue that indifference, rather than hostility, should mark the negative pole of that dimension. As will be shown next, the neighboring octants are also defined by notions of distance and detachment.

Diagonal Axes

The octants where the Extraversion axis is located [80, 39] are labeled "Aloof-Introverted" and "Extraverted-Gregarious" [138].

The former is associated with terms like "distant", "unsociable", "antisocial" and "introverted" in the IAS-R [138]. Its counterpart is defined by words like "cheerful", "friendly", "enthusiastic", "outgoing" and "extraverted".

In the IPIP-IPC questionnaire [79], Introversion is mapped to being "quiet around strangers", "a very private person", "[not] talk[ing] a lot" and "hav[ing] little to say". Extraversion, in turn, is equated to "feel[ing] comfortable around people", "start[ing] conversations", "talk[ing] to a lot of different people at parties" and "lov[ing] large parties".

Finally, the last two octants are "Arrogant-Calculating" and "Unassuming-Generous" in the IAS-R [138]. This reference gives adjectives such as "cunning", "boastful", "cocky" and "sly" for the former pole, and terms such as "undemanding", "unargumentative" or "uncalculating" for the latter.

The IPIP-IPC [79] measures "Arrogant-Calculating" with the subscales "cut others to pieces", "contradict others", "snap at people" and "have a sharp tongue". The opposite pole is measured using the statements "tolerate a lot from others", "take things as they come", "think of others first" and "seldom stretch the truth". Unlike the adjectives used in the IAS-R, which focuses more on the intellectual "calculating" aspect, the sub-scales of the IPIP-IPC are more intuitively aligned with the definition of Agreeableness. This again supports the theory that this personality trait is associated with these two quadrants [80, 39].

3.2.4 Three-Dimensional Interpersonal Models

Although the Interpersonal Circumplex is well-established and widely used, there is evidence that two dimensions are insufficient for modeling relevant nuances of interpersonal behavior. Some researchers have proposed distinguishing between different aspects of Affiliation, while others have argued that Extraversion and Agreeableness are not the only personality traits to determine interpersonal attitudes.

Distance and Affect

There is some disagreement about the labeling of the horizontal axis. Horowitz et al. [55] propose that its negative pole should not be active hostility but rather indifference. According to them, social behaviors invite reactions on the same side of the affiliative or communal axis, but on the opposite side of the status or agency axis. For example, a person giving friendly advice expects the other to accept it warmly, whereas someone telling the other to leave them alone would expect that person to withdraw obediently. Furthermore, they explain that hostility equals anger which is caused by incompatible reactions, even if those are well-intentioned. One example would be a situation in which both parties insist on politely yielding to the other, which can be just as irritating as a power struggle. From these observations, the authors conclude that interpersonal responses are best explained by modeling the degree of separation or closeness.

Spencer-Oatey [125] compared numerous psychological sources and found that, while most authors agree on the general meaning of the interpersonal dimensions, few give a precise definition and there are many terms that are used synonymously, but have different semantical connotations. Specifically, she argued that "distance" should be differentiated from "affect", since the former is more closely related to the duration of a relationship or the frequency of interaction whereas the latter refers to the positive or negative evaluation

of that relationship. For example, co-workers who have known each other for years are not necessarily friends, but could be bitter rivals instead. Similarly, people can have a positive or negative disposition towards people they hardly care about, such as store clerks or strangers at the bus stop.

The Extraversion-Agreeableness-Neuroticism Sphere

The three dimensions proposed by Spencer-Oatey [125] call to mind the PAD temperament model, as there is an intuitive mapping between both spaces. Both have the notion of dominance and submissiveness as one defining axis. Pleasure and its equivalent Valence can be aligned with the positive or negative feelings in a relationship, while Arousal or Activation seems to match the intensity and degree of involvement in the interaction.

This intuition is supported by Saucier [117] whose model combines the Affective and the Interpersonal Circumplex into a near-spherical construct. Moreover, this model is based on three of the Big Five factors: *Extraversion*, *Agreeableness*, and *Emotional Stability*. As explained in section 3.2.3, the Interpersonal Circumplex can be represented using the former two factors, while according to Mehrabian [86], the latter is strongly related to both *Pleasure* and *Arousal*. Saucier clustered the descriptive terms based on the octants of the three circumplexes formed by these personality factors. Later in the same year, Hofstee et al. [53] mapped the ten circumplexes formed by each pairwise combination of the Big Five factors, confirming most of Saucier's clustering. These clusters are shown in figures 3.6, 3.7 and 3.8.

The Extraversion-Agreeableness circumplex corresponds to the Interpersonal Circumplex, as was explained in section 3.2.3. Except for the terms *unagressive* and *obliging*, which Saucier associated with the *low Status* octant, the results of Hofstee et al. are in line with Saucier's work [117, 53].

Saucier relates the Extraversion-Neuroticism circumplex to the Affective Circumplex. Again, the mapping found by Hofstee et al. is mostly in line with his clusters [117, 53]. However, three of the terms from Saucier's *high Neuroticism* cluster - *nervous*, *fearful* and *fretful* - ended up in the neighboring *low Valence* cluster. Two more terms - *high-strung* and *sedate* appear slightly shifted towards low valence.

According to Saucier's results, Agreeableness and Neuroticism were the third pair of factors that form a complete circumplex without empty octants. Only two terms, *temperamental* and *soft* are slightly misplaced in the mapping by Hofstee et al. [117, 53]. Unlike the other circumplexes, this one has no well-known counterpart. However, it shows tendencies for positive evaluation and cooperative attitudes on the horizontal axis, while the vertical axis seems to reflect the degree of emotional intensity.

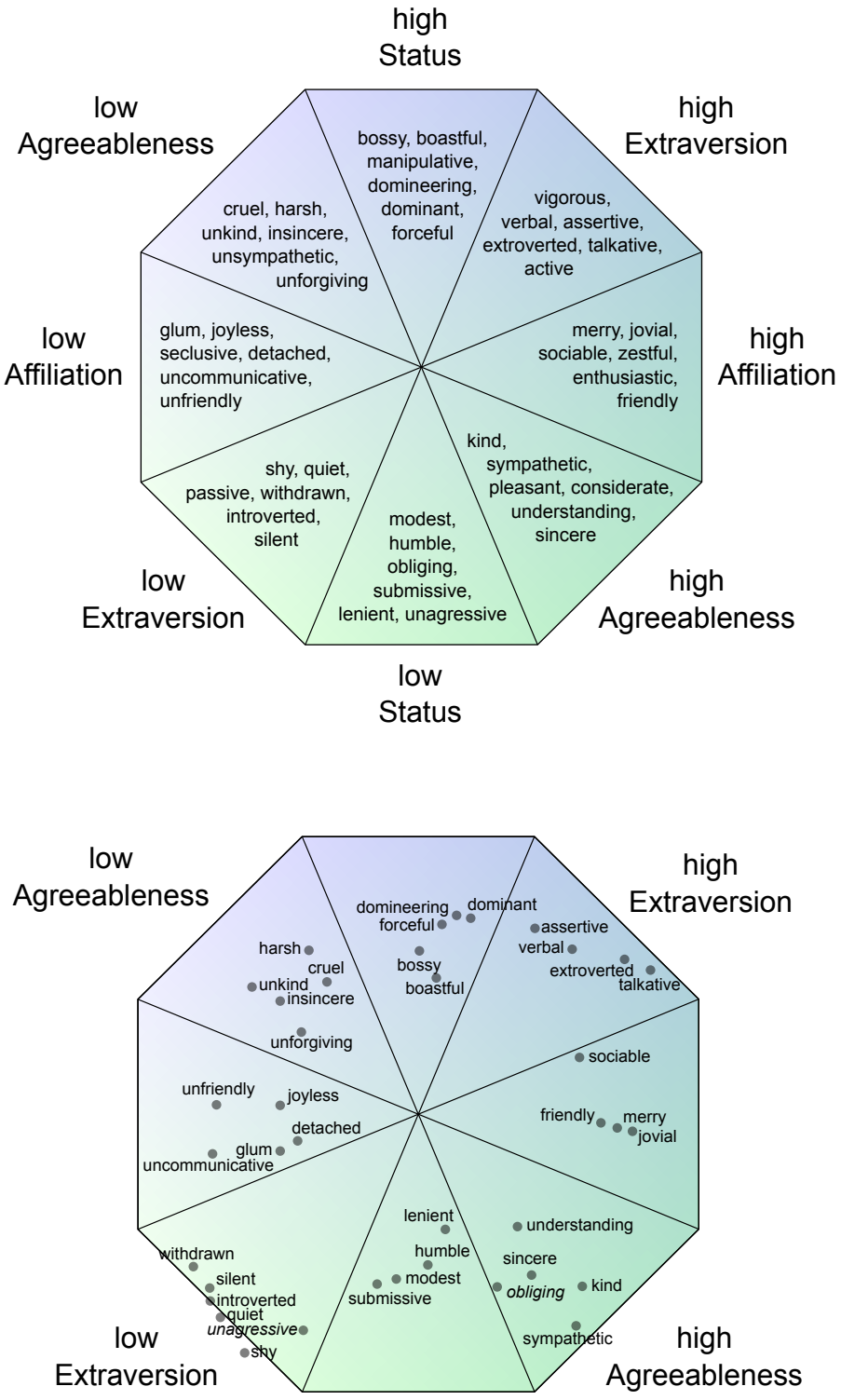


Figure 3.6: The Extraversion-Agreeableness circumplex. *Top:* According to Saucier [117]. *Bottom:* According to Hofstee et al. [53].

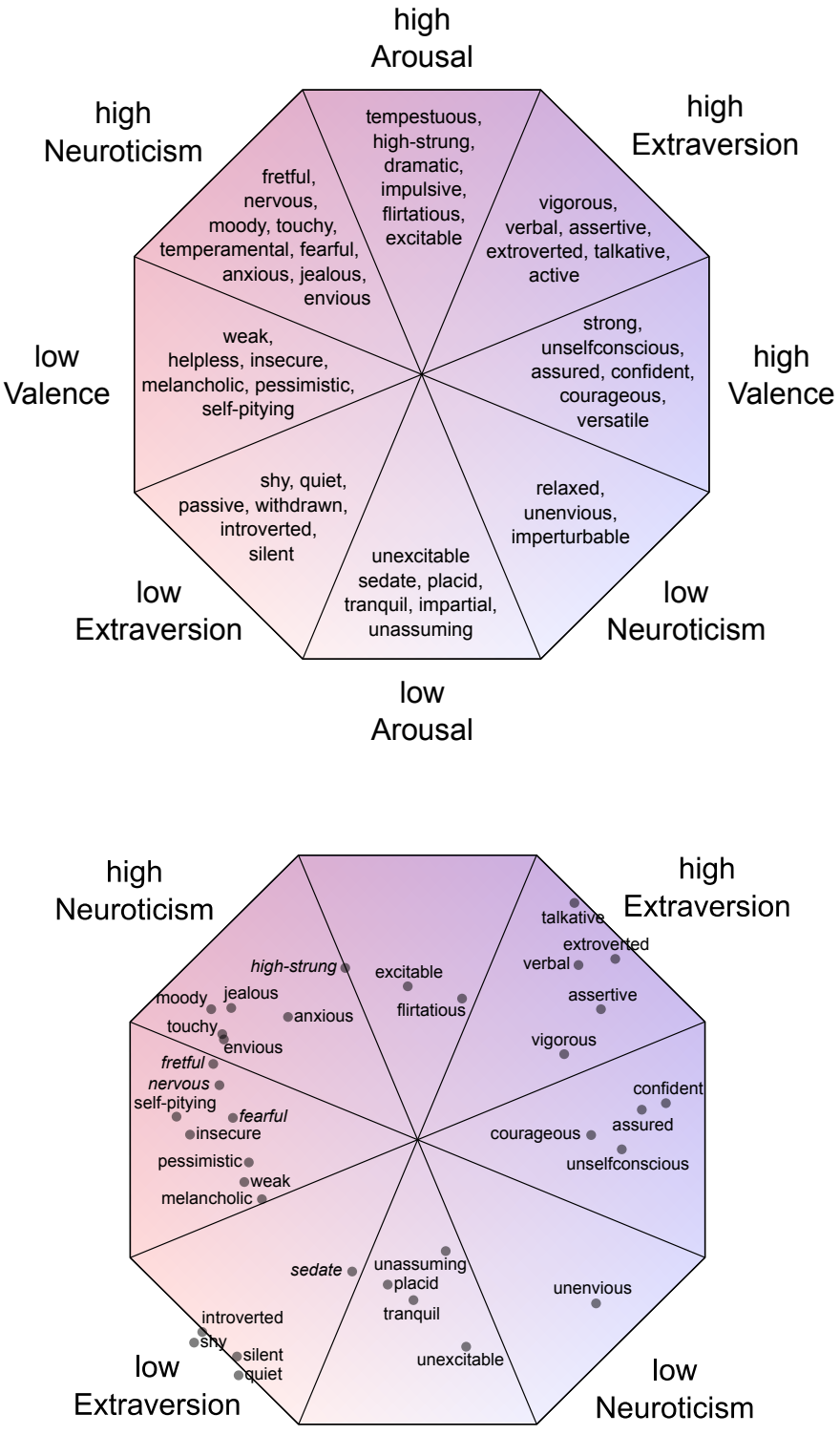


Figure 3.7: The Extraversion-Neuroticism circumplex. *Top:* According to Saucier [117]. *Bottom:* According to Hofstee et al. [53].

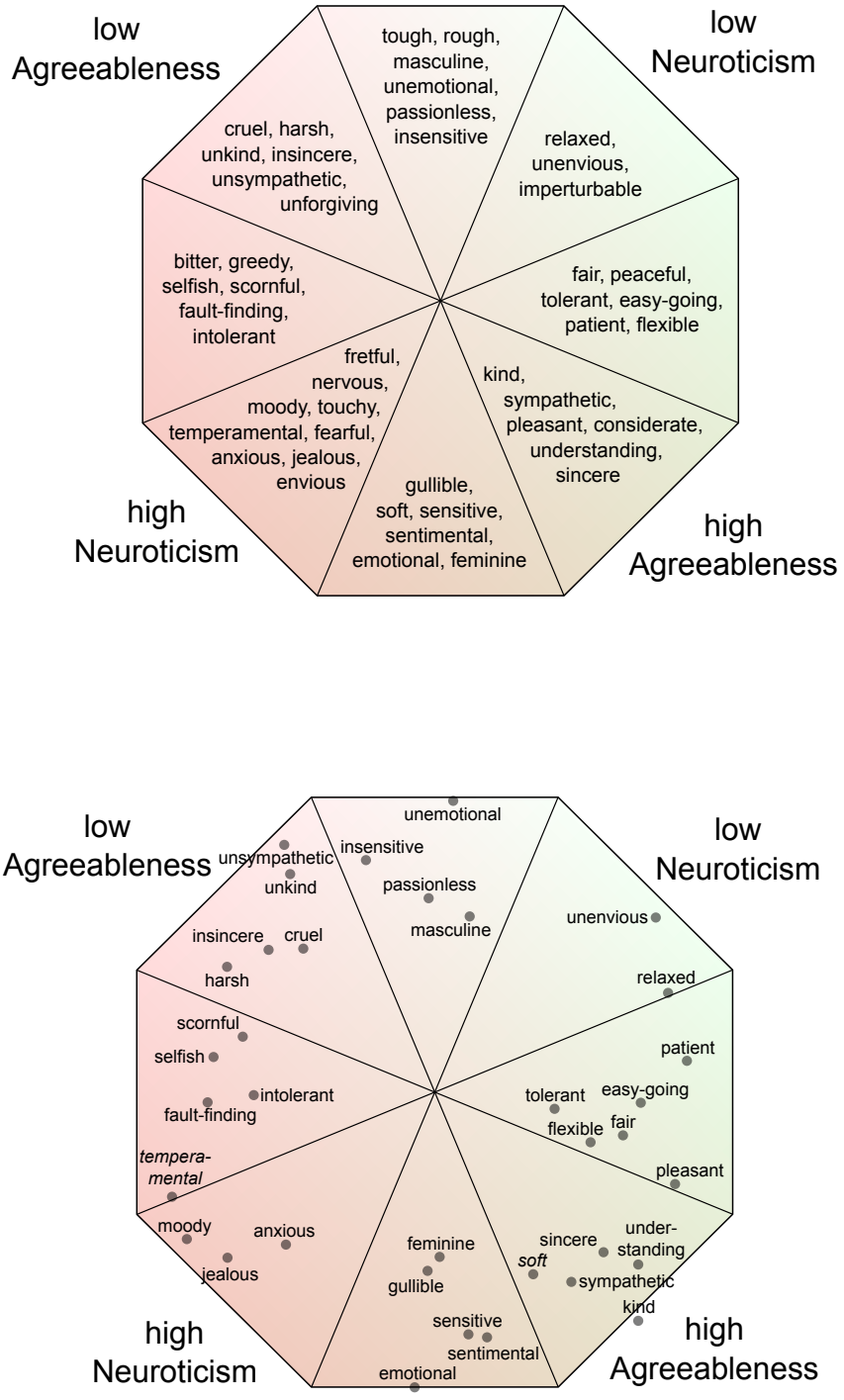


Figure 3.8: The Agreeableness-Neuroticism circumplex. *Top:* According to Saucier [117]. *Bottom:* According to Hofstee et al. [53].

Indeed, de Raad later suggested a third interpersonal factor related to emotionality or emotional expressiveness [37]. He analyzed 986 verbs related to interpersonal behavior and traits, thereby finding evidence for additional relevant dimensions. For instance, he found that the word clusters defining the two factors resembling *Dominance* and *Nurturance* deviated semantically from the established Interpersonal Circumplex. Additionally, his results hinted at axes such as *intimacy/comradery* versus *distance/withdrawal*, also referred to as *seeking* versus *avoiding contact*, and *agressive/emotional* versus *oratory*, which he called *demonstrative* behavior. Although the clusters he found do not exactly correspond to the Big Five factors, he suggests that the factor of Neuroticism respectively Emotional Stability "may be a good candidate to anchor parts of the latter cluster" [37].

Further support for this model can be found in literature which focuses on nonverbal communication. According to Argyle [7, p. 269], the traits Extraversion, Agreeableness, and Neuroticism seem to have the strongest influence on observable behaviors.

It should be noted that, while many behaviors are commonly associated with certain character traits, experiments tend to confirm participants' stereotypes rather than the actual connection to personality [71, p. 381-382]. One reason for this is that most established personality measures are subjective rather than objective. Self-report answers often differ from those given by others, and different relationships with the observers may trigger different behavior tendencies in the same person [11]. However, since the goal here is to create believable agent behavior, it is logical to use the traits that humans commonly associate with behavior tendencies.

Aligning the Circumplexes to PAD Space

Saucier expressed the Affective Circumplex using Extraversion and Neuroticism as orthogonal axes [117]. However, in 2001 Yik and Russell examined the mapping between that circumplex and the Big Five factors more closely and found that, although Affect was indeed most strongly related to Extraversion and Neuroticism, these factors were far from orthogonal [141]. According to their mappings, the Extraversion axis is rotated by 36° relative to the *Valence* axis while Neuroticism is rotated by 180°. These mappings can be seen in figure 3.9.

Similar results can be found in Mehrabian's mappings between the PAD temperament model and the Big Five factors [86].

According to Mehrabian [86], the PAD dimensions can be mapped to the Extraversion-Agreeableness-Neuroticism space as follows:

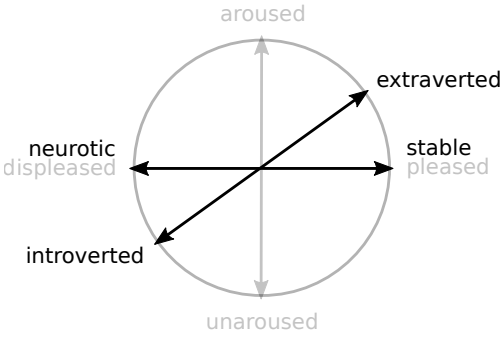


Figure 3.9: The locations of Extraversion and Neuroticism relative to the Affective Circumplex, according to Yik and Russell [141].

$$Pleasure = 0.21Extraversion + 0.59Agreeableness - 0.19Neuroticism$$
$$Arousal = 0.00Extraversion + 0.30Agreeableness + 0.57Neuroticism$$
$$Dominance = 0.60Extraversion - 0.32Agreeableness + 0.00Neuroticism$$

Figures 3.10, 3.11 and 3.12 show the relative placement of Extraversion, Agreeableness, and Neuroticism in PAD space.

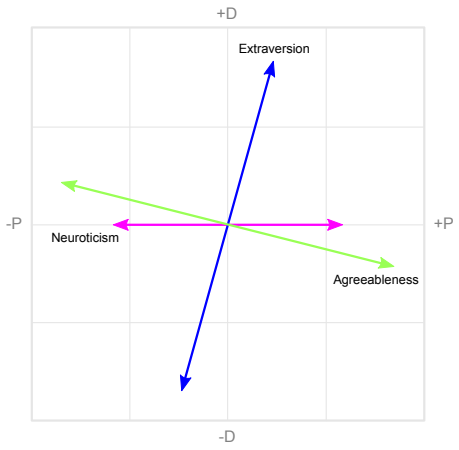


Figure 3.10: Location of Extraversion, Agreeableness and Neuroticism in the Pleasure-Dominance plane, according to Mehrabian [86].

As seen here, Extraversion and Agreeableness appear close to orthogonal (77.5°) on the Pleasure/Dominance plane. Extraversion is found at -27.2° relative to Dominance, and Agreeableness is located at -13.7° relative to Pleasure. These angles are roughly in line with previous findings about the Interpersonal Circumplex (see section 3.2.3), especially the mappings found by DeYoung et al. that hinted at smaller relative angles than 30° [39].

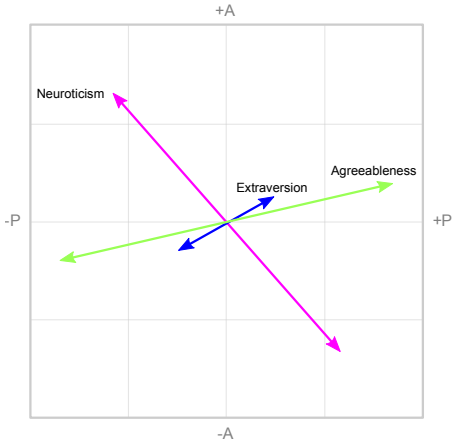


Figure 3.11: Location of Extraversion, Agreeableness and Neuroticism in the Pleasure-Arousal plane, according to Mehrabian [86].

As for the projection to the Pleasure-Arousal plane, Extraversion is close to the Pleasure axis, whereas Neuroticism is found in the middle of the negative Pleasure/positive Arousal quadrant (131.3°). The obtuse angle between the personality traits matches the findings of Yik and Russel [141], but the positions do not.

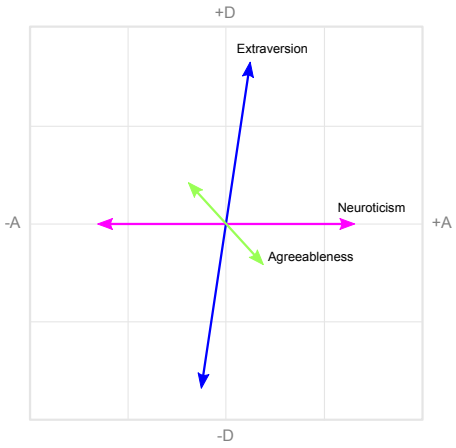


Figure 3.12: Location of Extraversion, Agreeableness and Neuroticism in the Arousal-Dominance plane, according to Mehrabian [86].

However, Extraversion and Neuroticism line up quite well with the axes of the Arousal-Dominance plane that does not correspond to a named circumplex. This directly contradicts Saucier's model that assumed that these personality factors represented the Affective Circumplex. According to Mehrabian's formulas, Agreeableness is a better match for the Pleasure axis than Extraversion, indicating that Agreeableness and Neuroticism form the Affective Circumplex.

There is obviously no perfect alignment between these different models. However, there is enough resemblance between them to help unify the literature on how the character traits in question relate to goals and actions.

3.3 Interaction Goals

Every action that an individual performs is an attempt to fulfill one or more of the individual's goals. For example, one might choose a specific manner for presenting an idea to one's boss to secure a promotion and consequently make more money. This, in turn, may be a subgoal required for meeting any number of long-term goals - improving one's social status, supporting friends or family members, enjoying a particular hobby, and so on.

Many of those goals, especially the short-term ones, involve other people besides the acting person. For example, they might need to collaborate on a task, trade information or physical goods with somebody, or deepen the social bond by showing interest and empathy for the other person.

To model the thought processes that lead to a specific behavior, it is first necessary to understand where the underlying goals come from and how they interact with each other. Therefore, this section will begin by reviewing classification schemes for different types of goals and how they can be aligned with the models described in section 3.2. Afterward, it will examine existing theories about how people deal with goal conflicts, both between their own and with those of others.

3.3.1 Goal Categorization

There have been different approaches to categorizing common human goals, such as grouping them according to semantic similarity, their importance to given individuals, or based on models such as the interpersonal circumplex [29]. This section will review several such works. In particular, it will focus on their relationship to the models for personality and interpersonal attitude explained in section 3.2.

General Categories

One possible way to classify a person's goals is by the source of the motivation. According to Barbuto and Scholl [15], there are five different motivation sources for human behavior:

- **Intrinsic Process Motivation:** fulfilling basic physiological needs and enjoying the action itself

- **Instrumental Motivation:** acting as a means to gain an advantage or a reward
- **External Self-concept-based Motivation:** conforming to the standards of the group in order to fulfill affiliative needs and gain status
- **Internal Self-concept-based Motivation:** conforming to one's own standards for the ideal self
- **Goal Internalization Motivation:** acting based on values shared with other people

Several of these sources appear relevant for turn-taking behavior. For example, one of the items the authors use to assess external self-concept motivation says: "It is important to me that others approve of my behavior". This hints at the need to conform to politeness rules. Said subscale also includes items concerning the need to be recognized for one's successes, or to make many friends in life.

In a similar vein, internal self-concept motivation is assessed by items such as "I try to make sure that my decisions are consistent with my personal standards of behavior" or "I like to do things which give me a sense of personal achievement". The former could be linked to a person's desire to be polite without external pressure, just because they believe in respecting other people or see themselves as a humble, unassuming person.

Finally, there is the intrinsic process motivation, which could be linked to enjoying the act of speaking itself - or conversely, to being unwilling to talk about unpleasant topics or to people one dislikes. This motivation source is assessed using items such as "I only like to do things that are fun" and "if I didn't enjoy doing my job at work I would leave."

While this categorization scheme already provides a useful guideline, it is not sufficient for relating specific goals to a structured model of interpersonal attitude or personality. Notably, there is no further distinction between status- and affiliation-related goals, which are all filed under the label of external self-concept motivation.

In 2001, Chulef, Read, and Walsh [29] employed a bottom-up clustering approach to better understand human goals. After identifying a suitable list of 135 goals, they had three diverse groups of laymen (173 people in total) sort those based on shared themes or topics. The subjects were free to create and label up to 30 categories as they saw fit. This way, the researchers intended to find the semantic concept structure as understood by the general population.

Their analysis revealed three prominent clusters: One covering to an individual's immediate family and sexual and romantic relationships, one concerning their interaction with friends and others in general, and one focusing

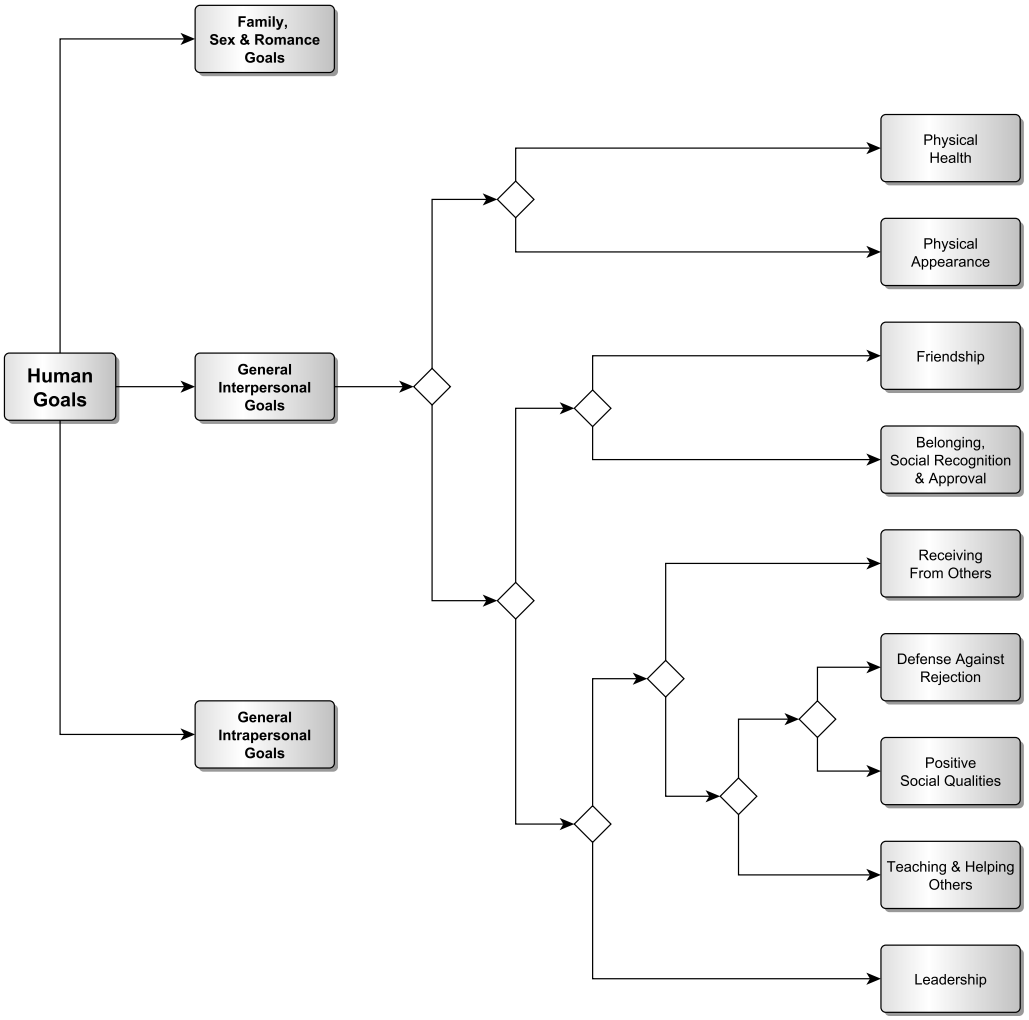


Figure 3.13: A subset of the goal taxonomy by Chulef et al. [29], focusing on interpersonal motives.

on themselves. The two latter goal clusters were labeled as *interpersonal* and *intrapersonal*, respectively. The subcategories that make up the interpersonal cluster are shown in figure 3.13, while table 3.1 elaborates on the specific goals sorted into said subcategories.

Besides physical appearance and health, which form their own subcluster, the interpersonal cluster includes goals that are intuitively related to the Interpersonal Circumplex. The subcategories hint at different levels of Status and Affiliation, or of Extraversion and Agreeableness (see section 3.2).

Leadership goals form a separate subcluster that mostly ignores the affiliation dimension and focuses on high status, with goals such as ”being better than others” and ”having control over others”. There are no indications of friendliness or hostility.

Category	**Associated Goals**
Physical Appearance	being clean/neat
	carrying oneself well/looking distinguished
	looking young
	being good-looking
	keeping up with fashion
Physical Health	looking physically fit
	maintaining a healthy weight/eating healthy food
	being physically active/exercising regularly
	being physically fit/in good physical condition
	having physical ability/agility
Belonging, Social Recognition and Approval	being in the center of things/popular
	being socially attractive/exciting/fascinating/impressing others
	being admired/recognized by others
	having a rich/active social life
	amusing/entertaining others
	knowing and being on familiar terms with many others
	belonging to/feeling like a part of social groups
	being likeable/making friends/drawing others near
Friendship	sharing feelings with close friends
	having friends/close companionship
	being affectionate towards others
Receiving from others	being taken care of
	having a mentor/someone to guide them
	having others to rely on
	receiving support from others on projects one believes in
Defense against Rejection	avoiding rejection by others
	defending oneself against others' criticism or attacks
Positive Social Qualities	being respected by others
	having others' trust
	being honest/loyal/respectful/courteous/considerate with others
Teaching and Helping Others	setting good examples
	helping others/cooperating/giving support
	developing others/teaching/sharing knowledge
	being in control of the environment
Leadership	being better than/beating others
	influencing/persuading others
	being in a position to make decisions for others
	having control over others
	being a leader

Table 3.1: Semantic categories of goals concerning general interpersonal relationships, according to Chulef et al. [29].

Its neighboring cluster covers providing and receiving support. It consists of the categories *Receiving from Others*, which indicates a submissive or dependant disposition, of *Teaching and Helping Others*, which indicates dominance, as well as *Defense against Rejection* and *Positive Social Qualities* which contain goals that fall anywhere on the status axis.

Finally, the rest of the interpersonal cluster is related to bonding with others. While the *Friendship* category is focused on being emotionally close, indicating high affiliation, the *Belonging, Social Recognition, and Approval* category also includes goals such as "being admired by others" and "impressing others" which indicate not only belonging but also high status within the group.

One shortcoming of this taxonomy is the lack of semantic labels for the intermediate clusters. To improve on this, Talevich et al. [129] repeated the study in 2017 with 489 naive participants and a revised list of 161 motives. Figure 3.14 shows an excerpt of the resulting structure.

Again, there are three primary clusters at the top level. The first one, labeled *Meaning*, concerns moral or religious values and ideals that the individual holds, as well as self-fulfillment and openness to experience. It covers a major part of the *intrapersonal* cluster described by Chulef et al. [29]. The second and third clusters are labeled as *Communion* and *Agency*. The former resembles the *interpersonal* cluster in the older work [29], comprising subclusters related to physical health and appearance, social bonds, and leadership. This time, however, goals related to family and sexual or romantic relationships are part of the same top-level cluster. The *Agency* cluster finally represents the remaining part of the aforementioned *intrapersonal* cluster [29], focusing on personal ambition, competence, and occupational success.

Attitude-based Goals

As evidenced by these general categorizations, numerous goals and motivations are directly connected to interpersonal relationships. It is therefore advisable to have a closer look at how they can be linked to the Interpersonal Circumplex.

Note that the terminology used by Talevich et al. [129] conflicts with earlier sources which use *Agency* and *Communion* as synonyms for *Status* and *Affiliation* [55]. The authors explain this by referring to an alternative distinction between *Agency* and *Communion*, namely the one between community-oriented and self-oriented individuals. Based on that distinction, they argue that having control over others and besting one's competition are inherently interpersonal goals. In contrast, goals related to general autonomy and mastery over one's own skills do not require any interpersonal relationship.

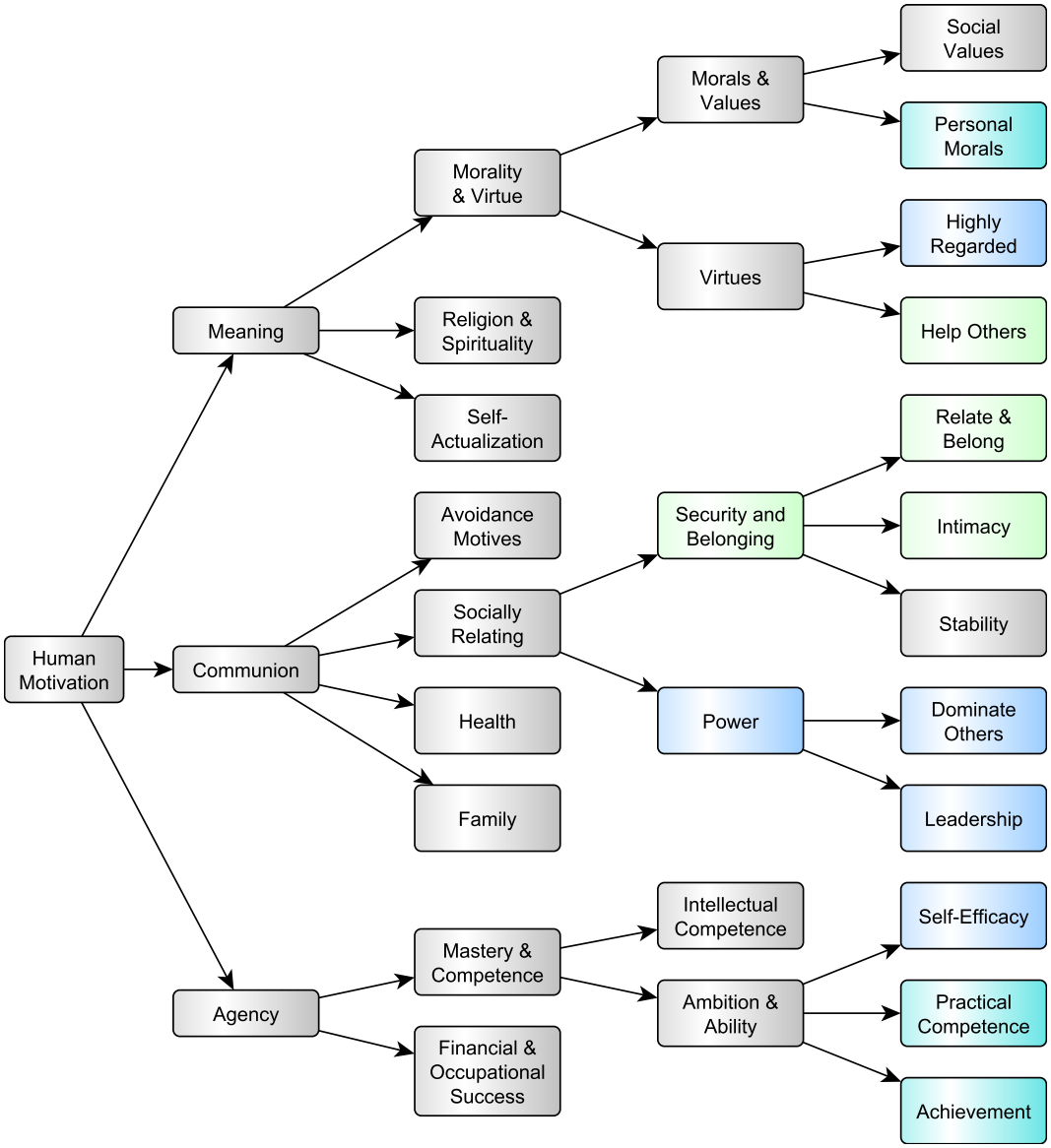


Figure 3.14: An excerpt of the goal taxonomy by Talevich et al. [129], focusing on social relationship goals. Green marks the nodes associated with the affiliation dimension, whereas blue marks those associated with status.

Before looking at the finer structure of the *Communion* cluster, two more categories for interpersonal behavior need to be explained. According to the *Politeness Theory* by Brown and Levinson, humans have two fundamental needs that they try to satisfy in social interactions [20, p. 61-62]. Those are tied to maintaining a person's public self-image, the so-called *face*. Brown and Levinson further distinguish between two aspects of said *face* and the associated *face wants*, which eventually lead to specific forms of politeness [20, p. 70].

- **Negative Face:** A person desires autonomy in their own actions. They want to be unimpeded by others and claim what is rightfully theirs. *Negative politeness* strategies focus on minimizing obligations for the other person or apologizing for interference.
- **Positive Face:** A person wants to feel appreciated and know that others approve of their goals. Therefore, *positive politeness* strategies involve expressing one's liking of the other person or treating them as a member of one's own group.

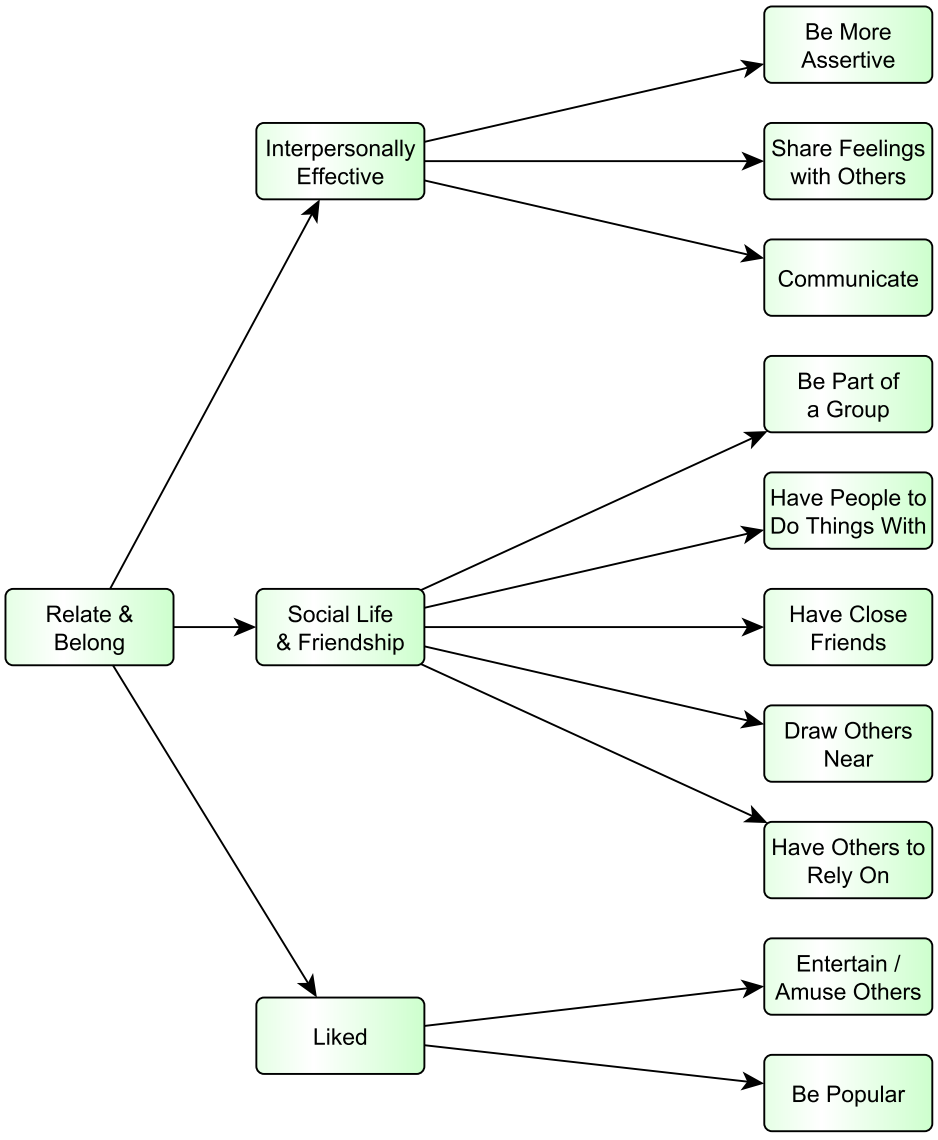


Figure 3.15: Subset of the taxonomy by Talevich et al. [129], focusing on affiliation-oriented motives.

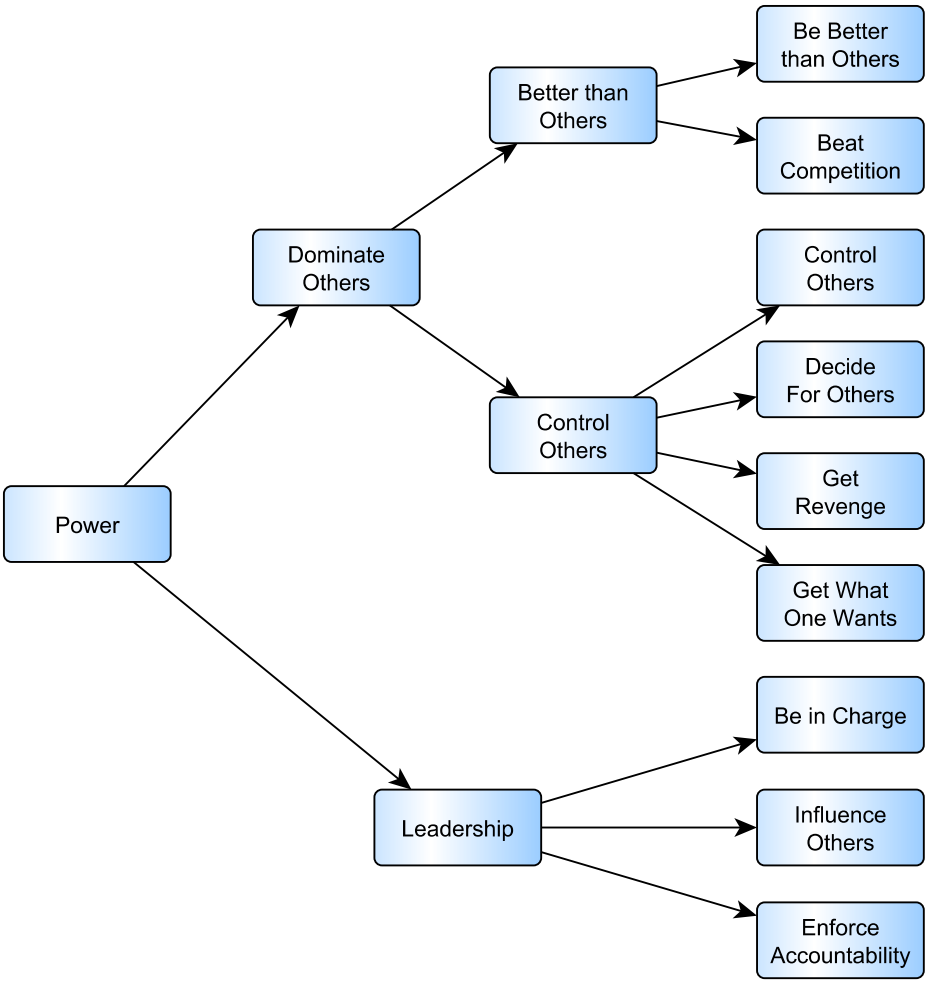


Figure 3.16: Subset of the taxonomy by Talevich et al. [129], focusing on status-oriented motives.

When examining politeness strategies used by socially anxious people, Oakman et al. [97] also discussed the relationship between Politeness Theory and the Interpersonal Circumplex. They argue that positive politeness has an intuitive mapping to the affiliation axis, whereas negative politeness can be equated to submission. Group membership plays a major part in positive politeness strategies. The degree of closeness or separation between people also is a defining aspect of the Affiliation dimension (see sections 3.2.3 and 3.2.4). Status, on the other hand, is related to drawing attention to oneself and imposing one's own will on others. These behaviors conflict with the other person's self-determination and threaten their negative face.

Looking back at the taxonomy by Talevich et al. [129], the *Communion* cluster is split into two subclusters labeled *Security and Belonging* respectively *Power*. *Security and Belonging* further consists of the subclusters *Intimacy*

and *Relate and Belong*, with the latter being more easily applicable to general human relationships.

Figure 3.15 shows the specific goals associated with *Relate and Belong*. There are three subclusters labeled as *Interpersonally Effective*, *Social Life and Friendship*, and *Liked*. The latter has the most straightforward mapping to positive face wants. The goals to *entertain or amuse others* and to *be popular* explicitly describe the desire to be approved of. The *Social Life and Friendship* cluster mainly contains goals related to group membership, representing affiliative needs. The connection to positive face wants is less explicit. Nevertheless, these can be found in the related politeness strategies which entail treating the other person as a member of one's own group. However, the third cluster - *Interpersonally Effective* - appears to cover a blend of high status and high affiliation. The goal *Be More Assertive* can be mapped to the status dimension, whereas *Share Feelings With Others* and *Express Oneself/Communicate* reflect the expressivity that is commonly associated with Extraversion.

As for the *Power* cluster, its goals focus on leadership, controlling others, or being better than them. These not only match the negative face want of autonomy but go beyond it by taking away other people's autonomy.

Personality-Based Goals

The mapping between the aforementioned motive clusters and the interpersonal circumplex implies that those same motives can also be associated with the Big Five personality traits of Extraversion and Agreeableness (see section 3.2.3).

In contrast, goals relating to the remaining three traits appear to be reflected in those motives which Chulef et al. [29] sorted into the *intrapersonal* cluster (see figure 3.17. Certain subclusters found by Talevich et al. [129] also match the definitions for those traits (see section 3.2.2). Specifically, both sources identify motives related to ambition and achievement, moral ideals, intellect and education, appreciating the arts and being creative, experiencing excitement, and seeking stability or avoiding negative situations.

In the newer taxonomy, both the *Agency* and *Meaning* clusters contain goals that appear related to Conscientiousness [129]. The subcluster *Ambition and Ability* (see figure 3.18) covers attributes like rationality, being organized, or perseverance. Other facets of Conscientiousness are reflected in the *Personal Morals* cluster that contains the goals "being honest", "being loyal", and "being an ethical person".

This same taxonomy also contains a cluster *Openness to Experience* that is mainly concerned with enjoying life, exploring, and appreciating beauty.

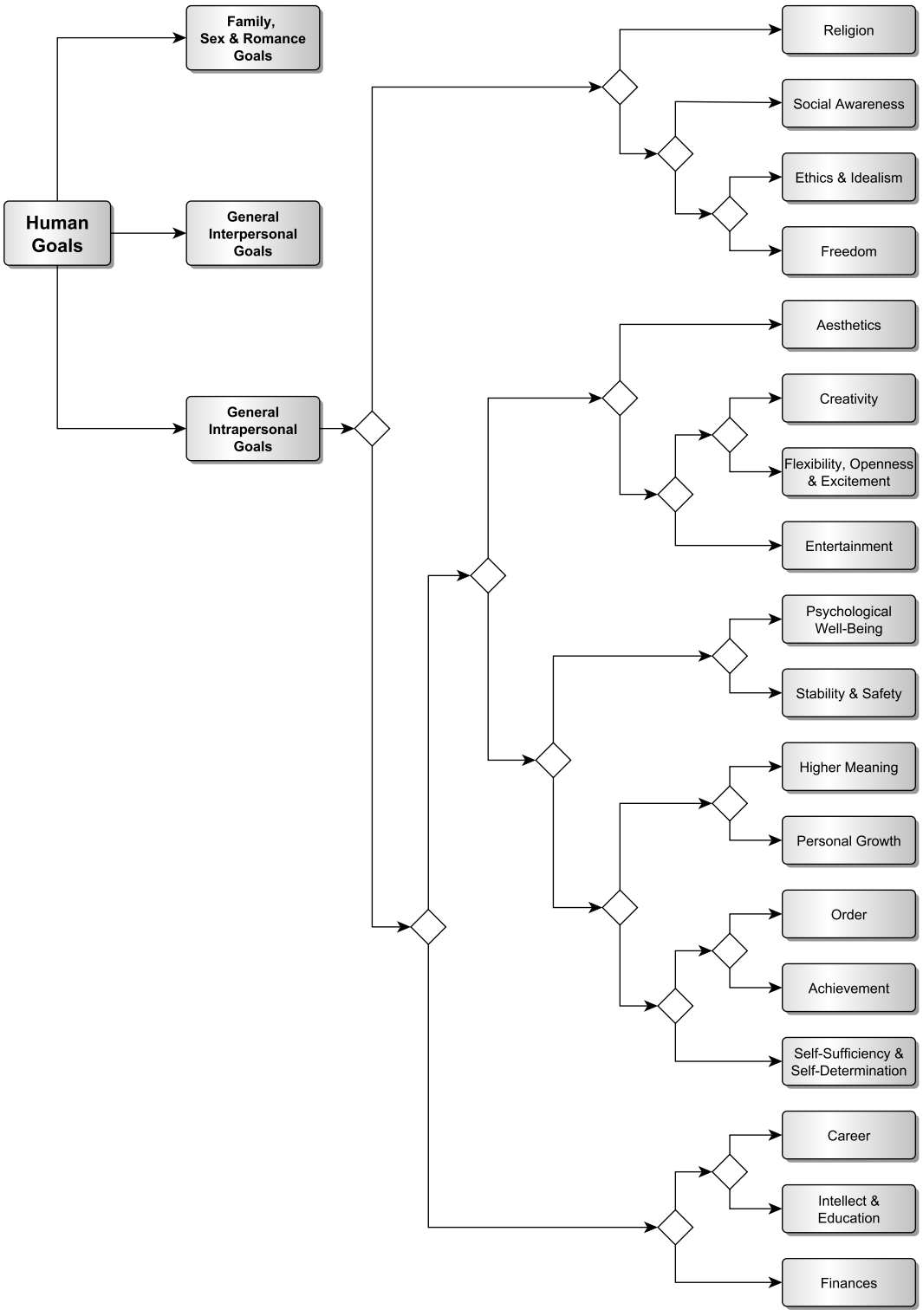


Figure 3.17: A subset of the goal taxonomy by Chulef et al. [29], focusing on intrapersonal motives.

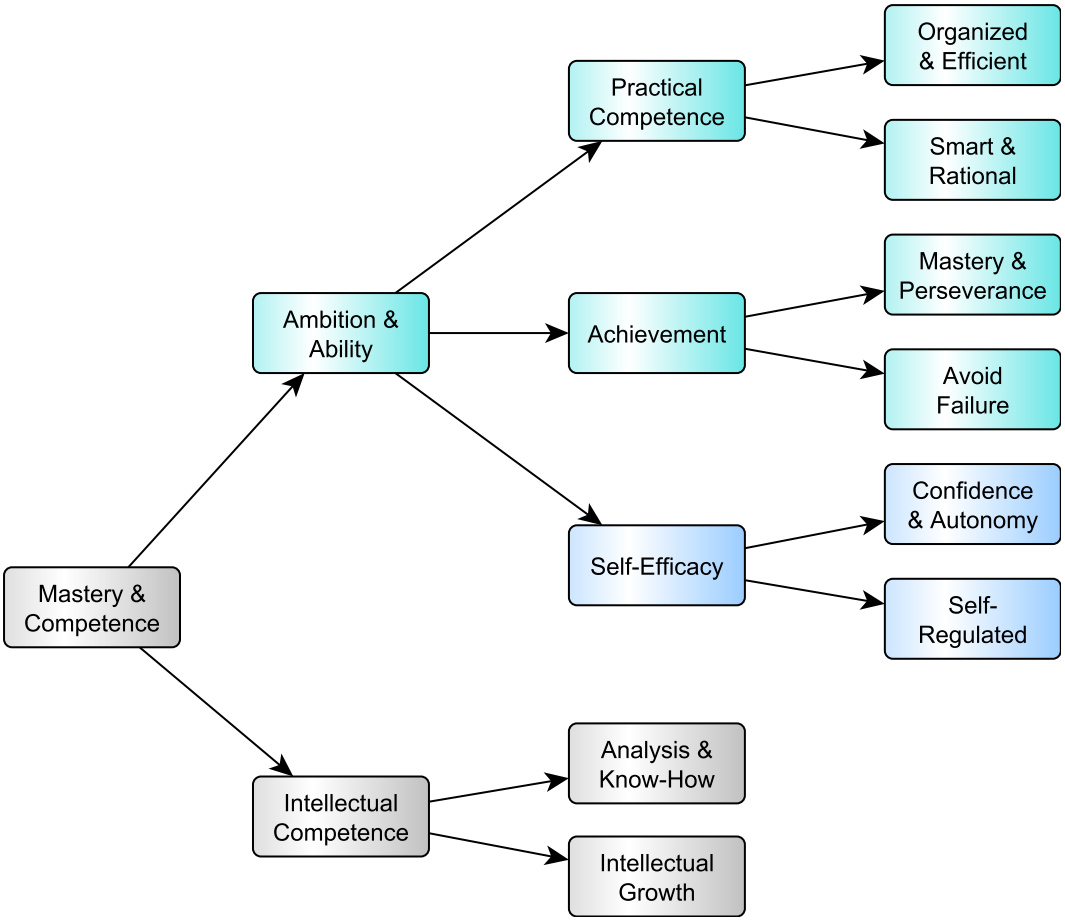


Figure 3.18: A subset of the goal taxonomy by Talevich et al. [129], focusing on ambition and competence.

Interestingly, while there are detailed clusters for Conscientiousness and Openness, both taxonomies contain only few goals related to Neuroticism. Recall that, back in section 3.2.4, the latter trait was suggested as the third one with a major influence on nonverbal behavior. The most relevant goals in the newer taxonomy can be found in the *Avoidance Motives* cluster (see figure 3.19). However, even there, they only take up a smaller subcluster next to others that are more closely related to Extraversion (e.g. "avoid socializing"), Agreeableness (e.g. "avoid conflict"), or Conscientiousness (e.g. "avoid effort").

One possible explanation could be that the counterpart of Neuroticism, *Emotional Stability*, implies satisfaction or indifference regarding the current situation and thus a lack of goals that need pursuing.

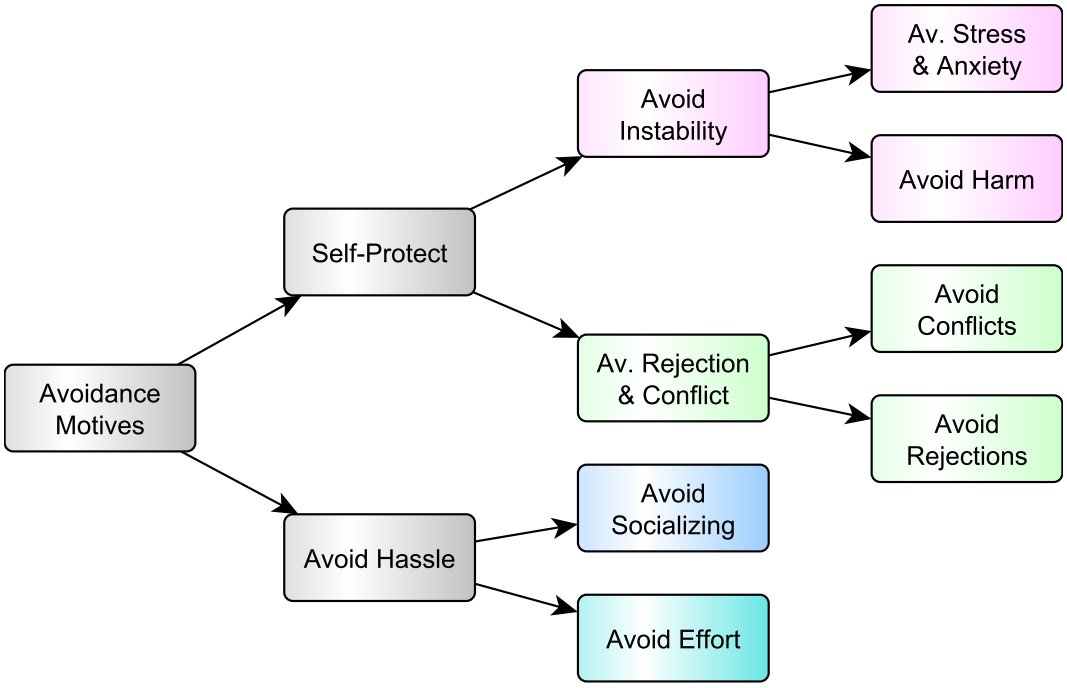


Figure 3.19: A subset of the goal taxonomy by Talevich et al. [129], focusing on avoidance motives.

Grounding

The goals mentioned so far are highly abstract and mainly concern long-term aspirations. Unfortunately, that makes them barely useful for determining the behavior of an artificial character. However, there is one salient topic in the context of interpersonal coordination that is commonly treated as neutral regarding personality or relationship.

Successful interaction - be it the exchange of information, joint action in a shared workspace, or the expression of empathy - requires the involved parties to agree on a certain set of beliefs and action plans. This shared knowledge is called the *common ground* and needs to be continuously maintained in a process known as *grounding* [32]. A closely related concept is *gaze cueing*, the phenomenon of one person looking at a target entity and others following their gaze [71, p. 299-300]. In other words, they are establishing common ground regarding the entities that are important at this moment.

Intuitively, there are several goals that can be related to the overall need for grounding. The most obvious one is the goal *"be understood correctly"*. It implies other goals, such as *"find out what the other party understood"* and, in case of a misunderstanding, *"clear up the misunderstanding"*. To avoid such misunderstandings in the first place, the fundamental goal *"establish common ground"* exists. This one, in turn, involves subgoals such as *"direct the other*

party's attention towards what is important" and *"confirm that the other party pays attention to what is important"*.

When moving to the social dimension of interaction, people tend to look more at those they like, which could be explained by a desire to see rewarding information such as signs of approval [71, p. 307]. It also allows for reading and exhibiting facial expressions, emotional or otherwise [8, p. 170][71, p. 298]. Taken together, these findings imply that people also establish common ground regarding each other's affective states. According to Ortony, Clore, and Collins, emotions arise from the appraisals of events, actions, and objects or people [98, p. 29]. Their model also assumes that said appraisal depends on, among other things, the "psychological closeness" between the observer and the directly affected person, as well as the degree to which the former likes the latter. Consequently, observing a specific emotion in a person can inform others of how they judge the situation at hand. When both parties know that the other person can see the emotional expression, the emotion in question becomes part of the common ground.

3.3.2 Goal Arbitration

People rarely have one single goal that they follow. Instead, there are often several goals involved which may or may not be in conflict with each other. A participant in a conversation often needs to decide between waiting for their turn and getting a word in edgewise. They might choose to break the silence, knowing that the other is just pausing to think and might feel offended by their impatience. Or they could have an urgent message that must not be delayed, no matter the cost.

Consequently, there is a need to assign priorities or weights to the different goals, enabling an interlocutor to select one at the expense of others.

Appraisal

One frequently-used model for simulating affective responses is the "OCC" model, named after its creators Ortony, Clore, and Collins [98]. The updated version, published in 2022, is also called "OCC2" to distinguish it from the original 1988 version.

The OCC model is built around appraising events, actions, and objects in relation to goals, standards, and tastes. The first two are relevant for turn-taking since an interaction process can be viewed as a series of both neutral events (e.g. something prevents the speaker from finishing) and explicit actions performed by a participant (e.g. the addressee barges in on the speaker's turn). As Ortony et al. point out, it depends on the individual observer whether they

only consider the outcome of an event or take the agency of another person into consideration.

Events, in particular, are evaluated based on what they mean for the goals that the observer wants to attain. Ortony et al. distinguish between more abstract "interest goals", such as wanting to lead a happy life, and more concrete "active pursuit goals" like wanting to buy a cup of coffee [98, p. 50]. They are assumed to form a hierarchy with less specific goals near the top and more clearly defined ones on lower levels [98, p. 54]. Low-level goals can *facilitate* or *hinder* the attainment of a high-level goal, or they can represent alternative paths towards it. These options can range from being *sufficient* to being strictly *necessary* for the high-level goal.

If an event has positive implications for such a goal, the observer may experience happiness or related emotions. Conversely, negative implications give rise to forms of distress. Taking the actions of another person into account can lead to gratitude or anger, respectively. A number of variables influences the intensity of the emotions. Of those covered in the OCC2 model , the following are most relevant for interpersonal coordination.

- **Desirability:** How beneficial an event is for oneself or the other affected person(s) [98, p. 60-65].
- **Psychological Proximity:** How close one feels to the event, the person(s) causing it or the one(s) affected by it. This proximity can refer to spatial, temporal, or social distance [98, p. 76-78].
- **Liking:** How friendly or hostile the relationship with the other person(s) is [98, p. 86]. Note that indifference towards the other(s) will not result in strong emotions because, as Ortony et al. explain it, people only respond emotionally when they care about something or someone [98, p. 49].
- **Praiseworthiness:** How well an action meets one's personal standards for behavior [98, p. 65].

Two of those variables are easily related to interpersonal attitude. "Liking" and "Psychological Proximity" call to mind the discussion about the *Affiliation* dimension and whether it should be split into two distinct concepts (see section 3.2.4). Looking at Spencer-Oatey's literature review [125], "Liking" matches the proposed dimension of "Affect" that is associated with positive or negative evaluations of the relationship. "Psychological Proximity", as mentioned above, includes the concept of social closeness. While Ortony et al. do not offer an explicit definition for social closeness, it makes intuitive sense to

use the same definition that Spencer-Oatey's review gives for "Closeness" (or its inverse, "Distance"). Examples given by Ortony et al., such as overhearing that a couple of either friends or strangers had lost their money due to a poor investment choice, indicate that people react more strongly to events that involve somebody whom they know well or consider part of their group.

Activation

The OCC2 model further assumes that not all goals are equally important at all times [98, p. 55]. Some disappear upon being attained, whereas others are transformed into goals for maintaining the new status. Others, such as the general interest goal "see people act in line with my own standards", are never fully attained, and the person having them is aware of that fact. Therefore, some goals remain in a person's mind even when there is no immediate plan or opportunity to work towards them.

Ortony et al. suggest that only parts of the goal hierarchy are *activated* at a given time, centered on the "active pursuit" goal that a person is currently focusing on. The superordinate goal may be used when explaining why the actively pursued goal is important in the first place. In contrast, those contributing to the success of the goal in focus are relevant when planning one's course of action. Goals that are further removed from the activated ones, such as long-term career plans, may be mentioned if someone has to go into detailed explanations of why they are doing something.

The consequence is that the OCC2 model only considers the activated subset of goals when appraising events or actions that lead to them. In the context of turn-taking, this means that although an interlocutor's long-term goal might be to build rapport with the other, they are more likely to focus on the conversation at hand and goals such as wanting to deliver a particular message or to hear the other participant's opinion.

Personality or temperament is commonly seen as the propensity for experiencing particular emotions in certain situations [87, 11]. Therefore, if the OCC model holds true and emotions arise from the evaluation of goal-related events, this hints at different goals being activated for different personalities in comparable contexts.

Prioritization

As detailed in section 3.3.1, Brown and Levinson [20] link politeness strategies to the assumption that humans strive for both individual autonomy and association with other people.

Spencer-Oatey [126] reviewed several works that challenge the idea of these goals being universal. However, her review also indicates that different cul-

tures may simply place different amounts of weight on these desires or have different expectations regarding the ideal levels of autonomy and association. For instance, eastern cultures tend to place more emphasis on association than autonomy whereas the western cultures tend to prioritize individual freedom [126].

According to Brown and Levinson [20], the wording or delivery of a message is determined by the degree to which the raw message content threatens the other person's positive or negative face. They suggested that this so-called "weightiness of the face threatening act" is composed of the relative social distance between the interlocutors, their difference in power, and the culture-dependent amount of pressure that this message puts on a person [20, p. 76].

One thing to keep in mind is that human decision-making is rarely optimal or logical [123]. Instead, it is skewed by different biases. People also tend to conflate aspects like the likelihood of obtaining a particular outcome with its desirability or approach it from the direction of which decision they would regret more if things were not to go according to plan [1, p. 41]. Consequently, experts in decision-making advocate for more structured approaches, like predicting quantifiable prospects based on domain knowledge and constructing explicit trade-off functions between different features of those [1, p. 98].

3.4 Coordination Mechanisms

Turn-taking is the process of coordinating who will start or stop speaking at which point in time so that the interlocutors achieve smooth transitions without collisions [41]. In a broader sense, it can be applied to silent collaboration in a shared workspace, purely social conversations, and hybrid forms such as discussing a complex issue with the help of reference objects.

Grounding is a vital ingredient for successful cooperation. As explained in section 3.3.1, all involved parties must share a certain amount of knowledge. This so-called *common ground* allows them to encode and decode messages, predict each other's actions, and adjust their own behaviors accordingly [32]. Therefore, participants are constantly establishing, maintaining, or repairing the shared beliefs about both the semantics of the participants' actions and the manner in which they are performed.

According to Clark and Brennan [32], each contribution consists of two phases. First, the sender *presents* a message, for example by speaking a sentence. Second, the receiver *accepts* this message by indicating how well they believe they understood the message. This acceptance can take the form of a short acknowledgment, usually called *back channel*, or a relevant response which forms a so-called *adjacency pair* with the accepted message.

Regardless of the interaction domain, the exchange of information is at the heart of it. Inferring the intention of others is important for planning one's actions while informing the interaction partner(s) of one's intentions is beneficial for working towards a common goal.

Therefore, paying attention to the other party is not only considered polite but also necessary for the interaction to succeed. Attention and information seeking are closely coupled, to the point that it is hard to say whether the behaviors associated with attentive listeners are consciously sent or just a by-product of closely observing the other party.

3.4.1 Information Seeking

Interacting parties generally wait for a sufficient amount of information before responding. Therefore, syntactic completeness tends to mark opportunities for changes in speaker and listener roles. Duncan listed syntactic completion as one of the signals for yielding the turn [41]. According to him, a complete grammatical phrase contains a subject and associated predicate.

As Goldberg explains it, a speaker not only has the right to transmit their message but also to receive a meaningful response to it [48]. This entails that the listener not only has the right to ask for clarification but even the duty to do so in order to honor the speaker's right to a proper reaction. Consequently, it is acceptable to interrupt the speaker if they do not provide sufficient, unambiguous information for an appropriate response.

In face-to-face communication, people rely on more than just audio information. Listeners are known to spend more time looking at the speaker than the other way around [8, p. 114][71, p. 299]. However, speakers do look at the listener when they try to gauge their understanding of what was said. Such gaze contact allows either party to observe the interlocutor's non-verbal reactions, such as a nod, a smile, or a confused frown [8, p. 121].

Furthermore, people look at relevant objects to plan the content of the conversation, for example, when studying a map while discussing a traveling route [10]. In that case, the conversation participants divide their visual attention between the interlocutor and the referenced object, with ratios depending on the complexity of the object and its relevance to the topic.

3.4.2 Attention Signals

Clark and Brennan explain that a speaker tries to speak at a time when the other person is "attending to, hearing, and trying to understand what he is saying" [32]. Consequently, sending and observing attention signals is essential for communication.

A person's focus of attention can often be inferred from the orientation of their head and eyes. This, in turn, allows a speaker to gauge whether the listener is paying attention to what they are saying, as evidenced by the latter monitoring the speaker's nonverbal behavior [71, p. 300] or shifting their gaze to referenced objects [10]. Thus, the act of gathering visual information itself becomes a piece of information that the interaction partner can gather.

Regarding turn management, humans are known to look at the interlocutor when they expect them to give a back channel comment or a full response to what has been said [8, p. 116]. This close link to information gathering suggests that humans may have ritualized this behavior and turned it into an explicit signal that tells the other person that the speaker is ready to receive information from them. In contrast, humans tend to avert their gaze when they do not wish to be interrupted.

Finally, gaze aversions can be explained as an attempt to avoid information when the cognitive load is high. In particular, humans are less likely to look at the other person while planning their sentence or searching for a specific word than they are during fluent speech or well-rehearsed phrases [70]. In other words, their attention is not on the interlocutor during those moments.

3.4.3 Feedback

Speakers continuously monitor the interlocutor's behavior for back channel communication that signals attention and understanding (see section 2.4.1). Besides that, the speaker also tends to be interested in how the other party evaluates the exchanged information. Listener feedback provides clues regarding their feelings about the topic at hand or how they relate to the interaction partner. Given their connection to an interlocutor's personality, those two aspects are particularly relevant for this thesis.

Affective State

As mentioned before, looking at an interaction partner's face is necessary for picking up their nonverbal signals. Consequently, people have been observed to look closely at those parts of a face where they expect characteristic deformations associated with certain emotions [71, p. 303]. These help both the speaker and listener understand how either party evaluates the information passed between them.

Turning one's face towards the other person can serve as an invitation to let them read those clues. Therefore, people tend to seek mutual gaze when they intend to transmit messages about their affective state and avoid it when they want to hide that information. For example, higher amounts of gaze have

been observed when physicians express empathy or when skilled liars try to appear innocent [71, p. 317]. Shame, on the other hand, leads to gaze aversion [71, p. 316].

Gaze aversion has also been linked to negative feelings [7, p. 166] as well as approach- versus avoidance-based emotions [71, p. 303]. The former calls to mind the "pleasure" axis of the PAD temperament model, which is used for describing both short-term emotions and long-term predispositions for experiencing them (see section 3.2.1). However, the emotions given as examples for approach (anger and joy) or avoidance (fear and sadness) are most clearly separated by said model's "dominance" axis (compare their placement in figures 3.2 and 3.3).

An alternative interpretation could be that the intensity of the emotional state controls the amount of gaze [7, p. 165], similar to the way gaze appears to reflect the intensity of the interpersonal relationship rather than one particular dimension [71, p. 309].

Social Relationship

Gaze has been linked to dominance as well as intimacy and liking [8, p. 170][71, p. 307].

Based on the attention mechanisms described earlier, making eye contact with another person indicates that one is interested in what that person is doing or expressing. Consequently, eye contact is a prerequisite for interaction, just as avoiding it is a means to prevent this interaction from taking place [71, p. 298]. Increased amounts of gaze serve to signal a wish for a closer, more intimate relationship, whereas people pay notably less attention to random strangers passing them by on the street, and completely ignoring others is seen as impolite and dehumanizing [8, p. 74][71, p. 300].

Continuous staring, however, is known to make people uncomfortable and trigger withdrawal or aggression [8, p. 92-93], possibly as a response to the perceived invasion of their privacy [71, p. 300]. Argyle and Dean [9] proposed that mutual gaze is an important component of intimacy. According to their *equilibrium theory*, people are therefore likely to reduce the amount of gaze when other factors (such as shorter distance or personal topics) increase the intimacy to an undesirable level.

Several studies report that people equate looking at another person with having a positive opinion of them [8, p. 58-61][71, p. 307]. A possible explanation is that people pay increased attention to friendly people because they expect to receive signs of approval from them [8, p. 61]. This would also be in line with the idea that the interpersonal dimension of affiliation can be mapped to the positive face wants from politeness theory (see section 3.3.1).

The fundamental principle of attention may also explain the hostile stare that seemingly conflicts with the previous phenomenon. According to Knapp et al. [71, p. 309], the intensity of the relationship, rather than its positive or negative evaluation, could cause this increased amount of gaze. To put it differently, a person glaring at another intends to interact with them to express their dislike or deliver a threat.

Averting one's gaze is typically associated with submissiveness or a lack of self-confidence [7, p. 167][71, p. 311]. On the other hand, it has been observed that lower-status group members pay close visual attention to their leader, especially while the latter is talking [7, p. 164-165]. High-status persons spend less time looking at lower-status speakers but gaze at listeners more frequently while speaking [7, p. 164-165]. The so-called *visual dominance ratio*, the amount of a person's speaker gaze divided by the amount of their listener gaze, has emerged as a more reliable indicator of status than the amount of gaze alone [71, p. 306-307].

A possible explanation for this difference in gaze patterns is that mutual understanding of the message is more important when it goes from a dominant group member to a submissive one. Consequently, both participants pay closer attention to feedback signals in this case [7, p. 164-165].

Besides gaze, the presence of gaps or overlaps in voice activity also carries information regarding the interlocutors' relationship. Rogers and Jones [113] examined the turn-taking behavior of human dyads. Each dyad consisted of one person with a high dominance score and one with a low score, and both interlocutors were of the same sex. The experimental results showed that the highly dominant interlocutor held the floor for 65% of the conversation. For every minute that the other person was speaking, they made an average of 2.71 attempts at interrupting the other, compared to 1.83 interruption attempts by their low-dominance counterparts.

The findings by Rogers and Jones support the common assumption that talking over the other person is a sign of dominance. Since they limit the other speaker's autonomy, interruptions and overlaps are generally considered impolite (see section 3.3.1 and the *face wants* defined by Brown and Levinson [20, p. 62]).

However, not all interruptions and overlaps indicate that one speaker dominates the conversation. Goldberg [48] proposes a spectrum from *power-oriented* interruptions to those that are *rapport-oriented*.

- **Power-oriented interruptions** typically change the direction of the conversation by introducing new, unrelated topics. Among these, assertions pose a greater threat to the speaker's *negative face* than questions

because, although the speaker does not control the topic anymore, they are still left in control over the answer.

- **Rapport-oriented interruptions** stay on topic. They tend to add information or express opinions about the speaker's statements, signaling that the interrupting party is interested in the topic at hand. This, in turn, can be seen as encouragement and an expression of shared goals, supporting the speaker's *positive face* rather than threatening it.

For the middle range of the spectrum, Goldberg mentions *competitive interruptions* and *quips*. The former type stays on topic but introduces power struggles through the participants trying to present their contribution as superior. The latter type depends on precise timing, leaving few options besides quickly seizing the conversational floor. They are usually light-hearted, which makes them less threatening, but they also show disrespect by interrupting the conversation flow, making fun of the speaker, or both.

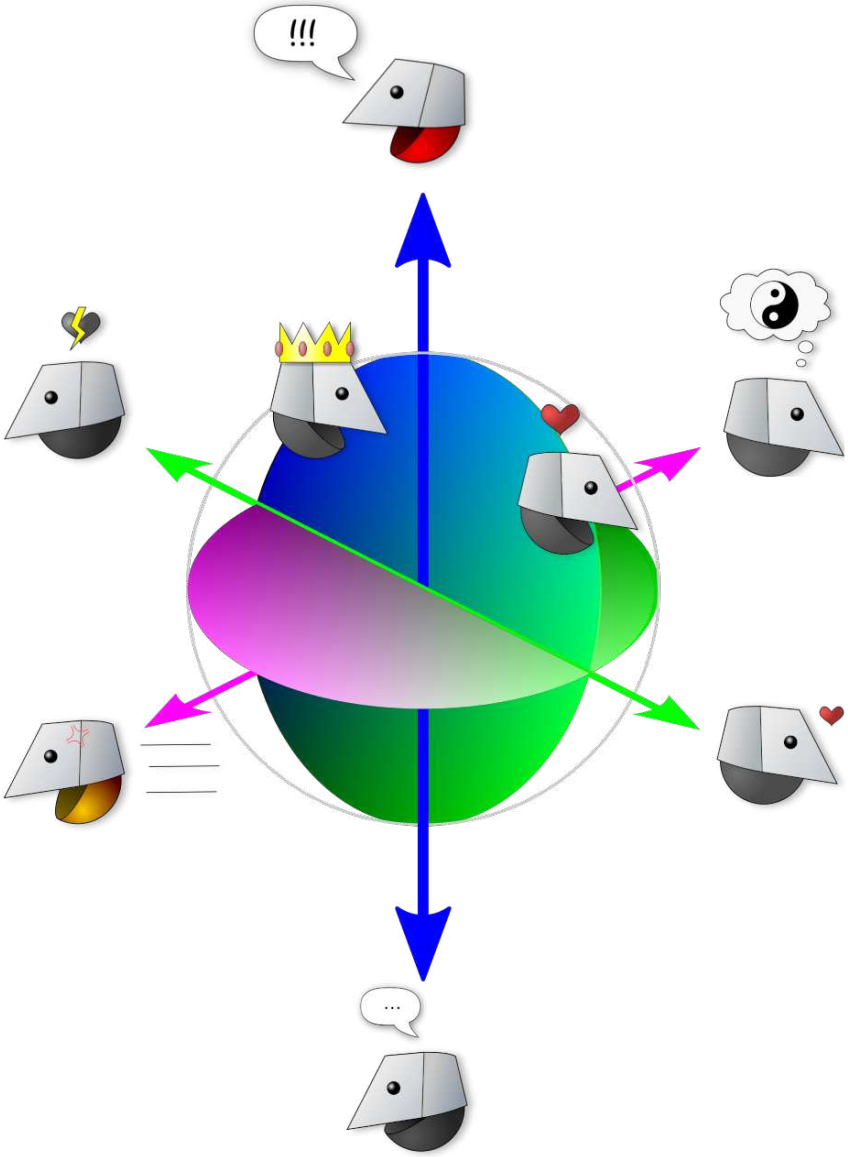
3.5 Conclusion

This chapter presented the psychological concepts and findings from which the agent's communicative behavior will be derived. In particular, these concern the models that will be used to define the agent's personality and attitudes, the goals that are relevant to the portrayed character, and the behaviors related to interpersonal coordination, such as turn-taking.

3.5.1 Personality and Interpersonal Attitude

There is evidence that personality is tightly linked to interpersonal attitude. Especially well-researched is the connection between the Interpersonal Circumplex and the traits of Extraversion and Agreeableness. This implies that a certain expression of said character traits produces a "default" attitude towards others, as indicated by the adjectives typically used to define them.

Regarding the interpersonal attitude, there are conflicting theories and definitions regarding its dimensions. Some authors consider splitting the Affiliation dimension into concepts such as the Closeness and the evaluation of the relationship. This would be in line with the OCC2 model of emotions [98], especially the idea that the intensity of social emotions depends on the psychological closeness to the affected person as well as the liking for them. More support comes from works that explore three-dimensional models. In particular, Neuroticism was suggested as a third trait with a strong influence on nonverbal behavior.



3.5.2 Interaction Goals

Information on specific communicative goals is hard to find. While there have been several attempts a building taxonomies of human goals, most of these are highly abstract or long-term goals. In the theory behind the OCC2 emotion model, these would be classified as "Interest Goals", whereas the modeling of conversational behavior would require specific "Active Pursuit" goals.

There are several clusters of goals that the taxonomy authors associated with interpersonal relationships. Many of these can be aligned with the Interpersonal Circumplex, both via the *Status* and *Affiliation* dimensions or the personality traits *Extraversion* and *Agreeableness*.

The remaining three of the Big Five personality traits appear linked to

goals concerning the self. Interestingly, there seem to be more goals linked to Conscientiousness and Openness than to Neuroticism, despite the latter being assumed to influence nonverbal behavior to a greater degree. One possible conclusion from this finding is that, while Conscientiousness and Openness might activate certain goals, Neuroticism appears to mainly influence a person's emotional response. If Neuroticism is connected to the perceived closeness between people (as indicated in section 3.2.4), then it might intensify emotions via the "psychological closeness" variable in the OCC2 model.

Speaking of activation, the OCC2 model also assumes that only a subset of a person's goals is activated at any given time. Several findings hint at different personalities assigning different weights or priorities to said goals, which in turn results in different emotional responses to the same situation.

3.5.3 Coordination Mechanisms

Interpersonal coordination is closely related to what is known as *Grounding*. Turn-taking, in particular, is a major part of *process grounding*, agreeing on who acts when. It relies heavily on gaze as a sign of attention that can be displayed or monitored in parallel to the actual message exchange. Apparently it is still unknown whether attention is signaled intentionally or is simply a byproduct of the search for information.

As for *content grounding*, the way a message is communicated provides additional clues regarding the relationship between interlocutors, as well as their current mood and emotions. Both are connected to the personalities that an observer assigns to the interacting parties.

Turn-taking behaviors, such as waiting patiently or barging in on the speaker, can be seen as a choice between respecting the other person's attention state or not. Accidental speech overlaps can be explained as a failure to establish common ground regarding said attention state. Consequently, attention will be a crucial ingredient in the behavior model that is developed in this thesis.

An interlocutor's attention is focused on what they consider important in that moment. This hints at a connection between the attention state that drives turn-taking and the personality-based weighting of interaction goals.

Chapter 4

Technical Background

4.1 Introduction

To bring psychological concepts into a machine, we need to translate them into mathematical formulas, numerical thresholds, and logical conditions. Fortunately, there are established computational models that aim to represent human decision-making and add a way to calculate the objectively best choice. The one chosen for this thesis is the *influence diagram*, a graph structure associating probabilistic outcomes with their *utility* for the decision maker's goals.

But a mathematical model on its own is of no practical use. To control an agent's behavior, it needs to be embedded in a software architecture that manages what the agent hears, sees, says, and does. At the University of Augsburg, many different graphical and robotic agents were available during this thesis but not all of them turned out to be suitable for exploring turn-taking behaviors. There were also different starting points for managing the interaction as a whole. However, none was readily usable with all agent types, and using different software frameworks was not practical in the long run. Choosing one was also made difficult by the sometimes conflicting, sometimes non-existent standards on which the various graphical and robotic agent platforms were built. Consequently, an important part of this thesis was figuring out the requirements for the agent control software and finding ways to extend existing frameworks.

This chapter gives an overview of the technical background that is necessary for understanding the presented research. The first section will look at decision theory, summarizing the mathematical principles behind it and explaining how they are represented in an influence diagram. After this, there

will be a section about semantic reasoning in ECAs and dialogue systems, followed by one about established agent frameworks and one about interaction management approaches. Another section will identify commonalities and differences of existing agent implementations before a summary will conclude the chapter.

4.2 Decision Theory

A decision-theoretic approach is used to make optimal choices in the face of uncertainty, by taking the probability of certain outcomes into account and weighing costs against benefits. Instead of going with their "gut feeling" or simple but flawed heuristics (as humans are prone to doing [123, 1]), decision-makers can systematically compare the expected consequences of different actions. For similar reasons, a decison-theoretic model avoids the problems that plague many machine learning approaches. Statistical models tend to rely on co-occurring patterns rather than true causal relationships. If the training data is skewed (for example, towards white men whose native language is English), the output will be skewed as well. In contrast, a decision-theoretic model not only focuses on known causal relationships but also forces the developer to consider all possible scenarios.

This section will explain the basics of probability, Bayesian inference, and utility-based reasoning. It is mostly based on the books "Learning Bayesian Networks" by Neapolitan [94] and "Foundations of Multiattribute Utility" by Abbas [1], which are recommended for diving deeper into this subject.

4.2.1 Probabilities

The first step in making an informed decision is to analyze which outcomes are certain, likely, or impossible. Furthermore, one needs to understand which context factors influence each other and which ones are independent.

Probability Distributions

A *probability distribution* describes the likelihood that a variable takes on a specific value. For example, flipping a coin usually has a 50% chance that the variable "top side" takes on the value "heads", and a certain place could have a 38% chance for the variable "weather" having the value "rain". In the context of turn-taking, there could be a 61% chance that an interlocutor starts to speak at a given time.

Unknown probabilities are often determined via the *relative frequency* of events. The variable of interest is sampled repeatedly, and the percentage of

samples showing a particular value is then assumed to represent the overall probability of this outcome. Figure 4.1 shows an example of determining a probability distribution this way. In this case, the variable "gaze" can take on four different values $x \in$ {"gaze at partner", "gaze up", "gaze to the side", "gaze down"}. The probability $P(gaze = x)$ is the number of matching observations divided by the total number of samples.

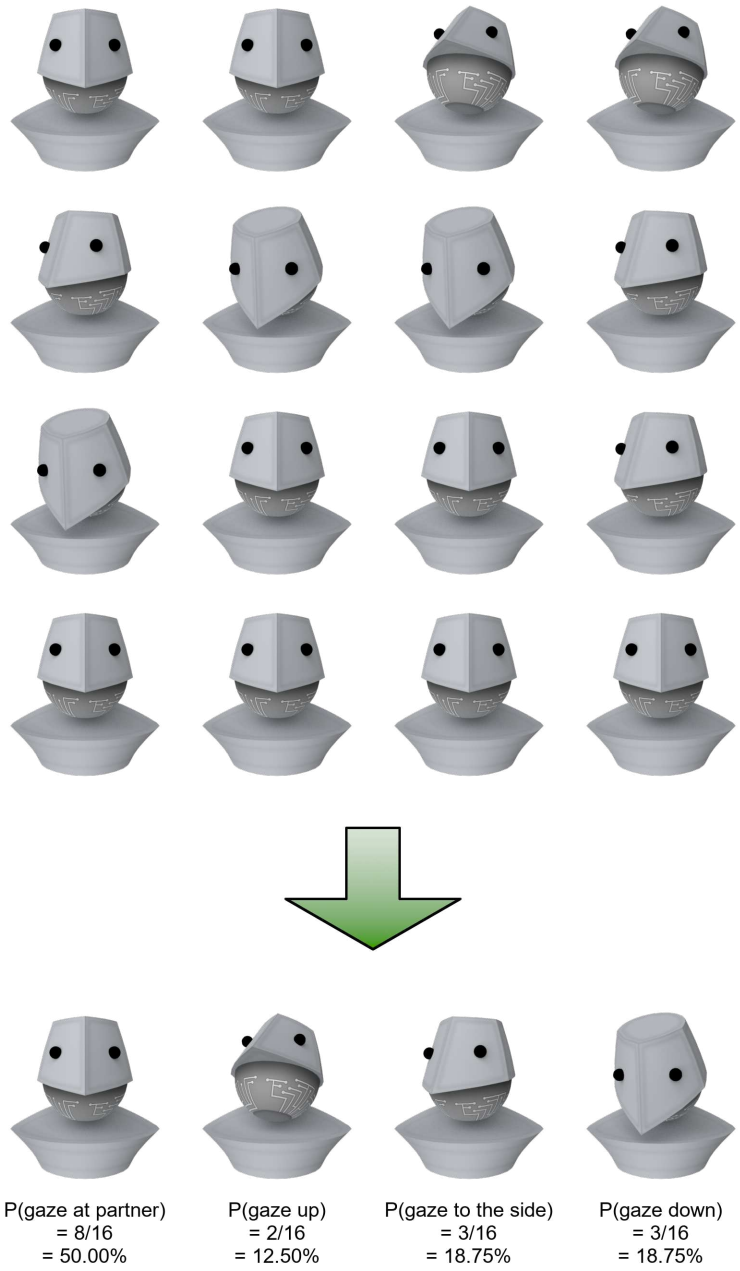


Figure 4.1: An example of using relative frequencies to determine the probability of observing a certain value for the variable "gaze".

Conditional Probabilities

Certain observations depend on each other, which means that the value one variable takes on changes the probability of observing a particular value for a different variable. For example, a person who is interested in the topic of the conversation is more likely to pay attention to the speaker than someone who is bored or distracted. Similarly, a listener looking at the speaker is more likely to be perceived as "interested" than someone looking out of the window - unless the speaker is talking about something that can be seen through the window.

Mathematically, this means that the probability P(look at speaker) is different when the variable "interest" has the outcome "interested" than when its outcome is "uninterested". For example, let's assume the following probabilities:

$$P(speaker|interested) = 0.8 \qquad P(speaker|uninterested) = 0.4$$
$$P(window|interested) = 0.2 \qquad P(window|uninterested) = 0.6$$

These dependencies also allow for inferring one observation from another, as is already hinted at in the statements above. For this purpose, however, we need to know the so-called *prior* probability distributions for the variables in this situation.

As explained before, this kind of information could be obtained by observing this listener during a large number of conversations, with a balanced selection of interesting and boring topics. Let's assume that they spent 70% of the observation time looking at the speaker. Let's also assume the speaker spent 60% of the time talking about topics that interested the listener. In the absence of other information, this gives us the following probabilities:

$$P(speaker) = 0.7 \qquad P(interested) = 0.6$$
$$P(window) = 0.3 \qquad P(uninterested) = 0.4$$

With that information, **Bayes' Theorem** can be used to calculate the probability that this listener is uninterested when they are looking at the window.

$$
\begin{aligned}
P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\
&\Downarrow \\
P(uninterested|window) &= \frac{P(window|uninterested)P(uninterested)}{P(window)} \\
&= \frac{0.6 * 0.4}{0.3} \\
&= 0.8
\end{aligned}
$$

Bayesian Networks

So-called *Bayesian networks* are an established graphical representation of such conditional probabilities. They consist of several *chance nodes* that each represent one variable along with the probability distribution for observing its possible values. They are connected in a *directed acyclic graph*, which means that the edges between the nodes point in a specific direction and that no loops lead back to earlier nodes in the connected chain or *path*. Nodes that come before a given node in that path are called *ancestors* whereas those that come afterward are called *descendents*. Immediate ancestors are called *parents*.

An edge from node A to node B means that there is a conditional dependency between the associated variables. Additionally, these connections must satisfy the *Markov Condition*, which means that the variable at any node is independent of those at its non-descendents when those at its parent nodes are given.

Figures 4.2 and 4.3 give an example of this independence. In the absence of further observations, the probability distribution of the variable "other gaze state" changes slightly when the outcome of its ancestor "other voice state" switches between "silent" and "speaking". However, if the outcome of its parent "other feedback need" is known, the outcome of "other voice state" becomes irrelevant, and the distribution of "other gaze state" remains the same.

4.2.2 Utilities

To choose the best strategy, one first needs to specify the consequences of each action. These can then be compared with what is needed to achieve

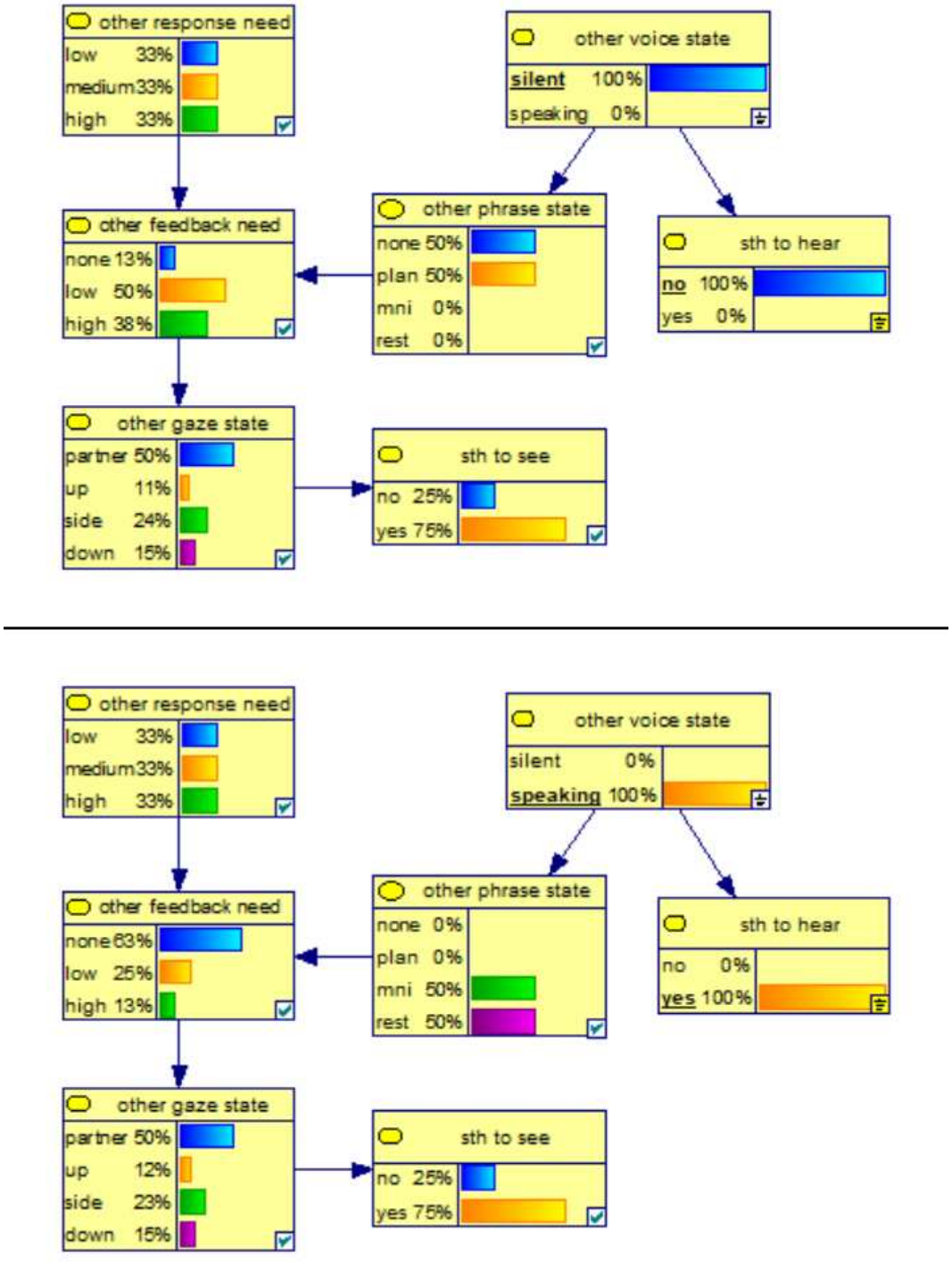


Figure 4.2: Ancestor "other voice state" influencing the probability distribution of "other gaze state".

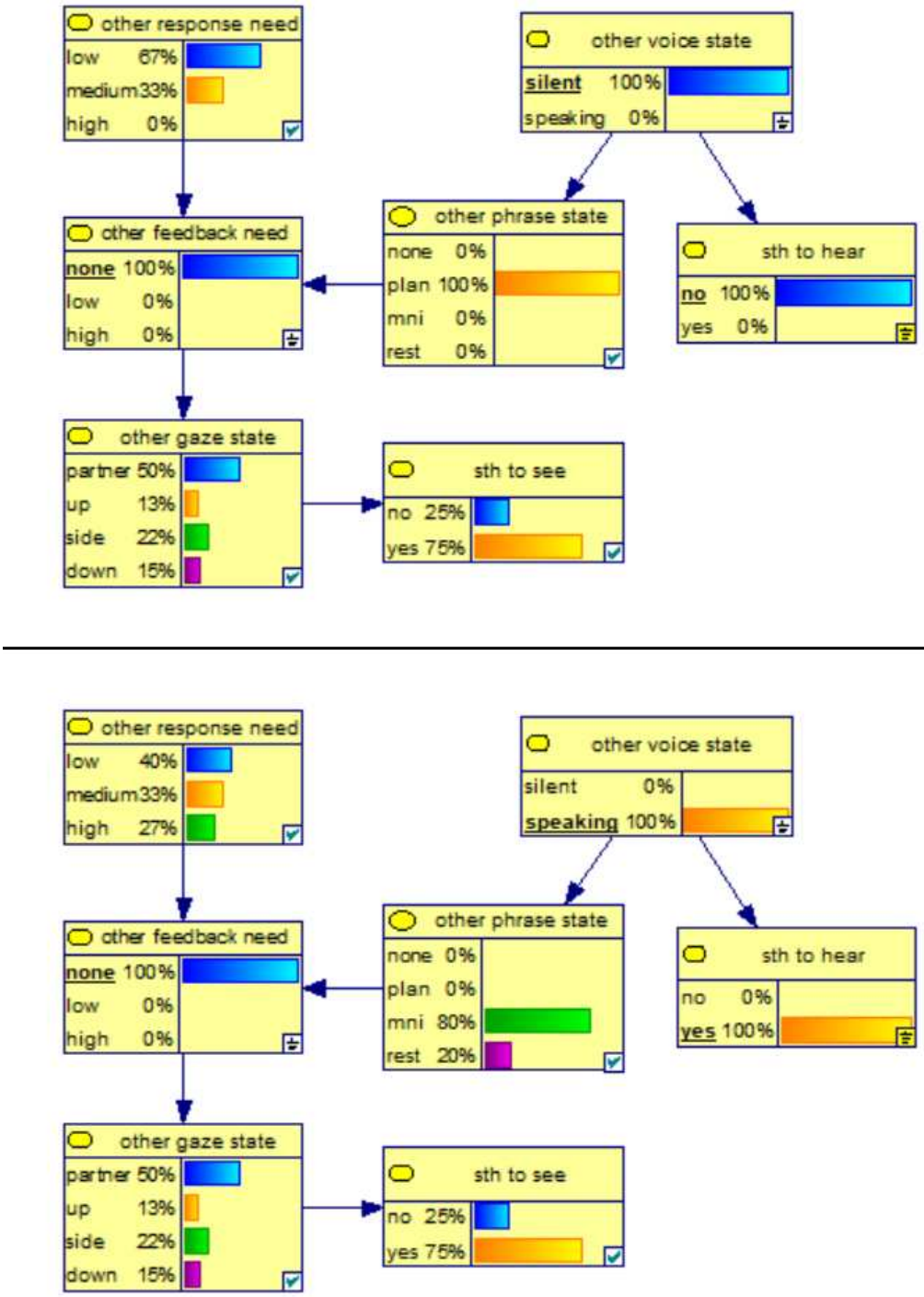


Figure 4.3: The fixed outcome for parent "other feedback need" preventing ancestor "other voice state" from influencing the probability distribution of "other gaze state".

a particular goal. Furthermore, weighting factors allow for prioritizing one goal over another. A conflict between different goals can then be resolved by selecting the action that is most likely to have the best consequences overall.

Prospects

According to Abbas [1, p. 4], a *prospect* is a possible state of the world after a decision was made. For example, when a person decides to start speaking, their voice will usually be audible after that point. Other factors beyond the decision maker's control can influence this world state as well. If another person is speaking at the same time, the former's voice may be drowned out, and their message may go unheard. However, if nobody else speaks, the same message will be heard very clearly.

Abbas further recommends using prospects that can be objectively measured or calculated from related variables, for example, using models from physics [1, p. 163]. In the context of turn-taking, such a prospect could be the duration of a conversation segment with overlapping speech. However, such objective measures are hard to find for social interaction. Personality traits, for instance, are still determined by having a person judge themselves or someone else with respect to related statements. The quality of a conversation could be determined by a fair distribution of speaking time, but participants are unlikely to know this distribution without reviewing a recording. They might be more interested in the amount of information that they gained or their success rate in expressing their own thoughts. Consequently, psychological models about event appraisal play a major role in determining the prospect of a turn-taking decision.

Utility for a Goal

A closely related but distinct concept is the *utility* of a decision outcome. The prospect is the same regardless of who is affected, but its usefulness for a particular person can vary greatly.

There are different approaches for mapping prospects to utilities. For example, if the prospect is expressed as the duration of a turn conflict in seconds, the utility could be the inverse of that duration to reflect that a person considers long conflicts damaging to their goals. Many examples in literature use money to represent the consequences of decisions, such as the cost of a repair [94, p. 241-246] or the outcome of an investment decision [94, p. 233-242]. For instance, Horvitz, Koch, and Apacible had office workers assign dollar amounts to different types of interruptions [56].

Alternatively, the utility can be a measure of indifference regarding different prospects. Abbas defines utility as the probability a person would want for

obtaining a more valuable prospect in exchange for a certain one, at the risk of getting a less valuable prospect if the gamble fails [1, p. 11]. To simplify this calculation, he recommends systematically mapping all possible prospects to a numerical value.

In the context of turn-taking, not speaking at all could be mapped to a utility of 0.0 regarding the goal of transmitting a message. If being the only speaker, and therefore heard clearly, was assigned the utility of 1.0, the outcome of talking at the same time as another person would have a utility between 0.0 and 1.0. The exact utility of the latter case may depend on a wide range of additional factors, such as the volume of the respective speakers' voices, or the degree to which the person in focus cares about being understood properly.

The *expected utility* of a chosen action is calculated as the sum of each prospect's utility weighted by the probability of obtaining that prospect after the action was performed.

$$EU(action_i) = \sum_{j=1}^{\#prospects} P(prospect_j|action_i)U(prospect_j)$$

The action that has the highest expected utility is then chosen because it is most likely to achieve the goal.

Multi-Attribute Utilities

Multi-attribute utility theory approaches complex decisions by breaking the alternative prospects down into a set of comparable attributes [123, 1].

There are different ways to combine these attributes into a prospect's overall value. One of the simplest methods is to add them in a weighted sum, with the more important attributes having a greater impact on the result [123]. However, this method is only appropriate for attributes that directly contribute to the value, as opposed to increasing the likelihood of getting certain values for the others [1]. Furthermore, it works best for attributes that can be measured on a clearly defined and meaningful scale [1]. Another aspect to consider is the *trade-off* between different attributes, the necessary increase in one attribute that would compensate for a decrease in another [1].

Regarding turn-taking goals, the utility of observing a speaker could be composed of attributes such as the amount of information they transmit verbally and nonverbally, respectively. Moderating factors could be the clarity with which the signals can be perceived, scaling the raw value of the available information. So, for example, the amount of received information via audio and visual channel would be modeled as follows:

$$
\begin{aligned}
& info_{received} \\
=& info_{received,audio} + info_{received,visual} \\
=& info_{sent,audio} * clarity_{audio} + info_{sent,visual} * clarity_{visual}
\end{aligned}
$$

Influence Diagrams

A Bayesian network can be augmented with nodes that represent decisions and the utilities of their consequences. Neapolitan [94, p. 253] depicts chance nodes as circles, decision nodes as rectangles, and utility nodes as diamond shapes. The GeNIe software[1] marks chance nodes with ellipses and utility nodes with hexagons.

Influence diagrams are mathematically equivalent to decision trees that lead to a different outcome depending on the choice one made [94, p. 233]. However, they are a more compact representation of the underlying decision process. They build on the Bayesian network's mechanisms for inferring the relevant probabilities and use conditional independencies to avoid redundant branches.

The expected utility of a decision is calculated by summing up the matching values at each utility node that depends on it. When there is more than one decision to make, they are evaluated in a fixed sequence, and the utilities for the possible combinations of choices are added up. More detailed explanations are provided by, for example, Neapolitan [94].

4.3 Semantic Reasoning

Some turn-taking patterns, such as looking at the listener at the end of one's turn, are independent of what is said. Others, however, depend on whether the semantic content was transmitted successfully, and the overall topic can influence whether observers perceive overlaps as status- or rapport-oriented [48]. Therefore, semantics cannot be ignored despite the goal of creating a domain-agnostic turn-taking model.

This section gives a brief overview of semantic reasoning in a computer science context.

[1] by BayesFusion, LLC, and available free of charge for academic teaching and research use at http://www.bayesfusion.com/

4.3.1 Belief-Desire-Intention Framework

One established framework for modeling agent behavior is based on the cognitive components *belief*, *desire*, and *intention*, commonly abbreviated as "BDI". Rao and Georgeff [105] described these components as follows.

- **Belief:** The information that the agent has about the current state of its surrounding environment. This world state is pieced together from numerous sensor updates and stored for reference.
- **Desire:** Goals that the agent needs to pursue, as well as their priorities and the trade-offs between them.
- **Intention:** The actions that the agent selected based on its belief. These are the actions that the agent will try to perform.

However, Rao and Georgeff also point out a major challenge in dynamic environments [105]. Since the world state can change during both the selection and execution of the action, the agent may need to abandon its intentions and select new ones if the original plan is no longer appropriate. However, replanning upon every new information would take time, allowing the world state to change again and exacerbating the problem. Therefore, the authors suggest that, depending on the application scenario, the agent might require different degrees of commitment to its intentions. A strong commitment would make the agent ignore new information for longer, whereas a weak commitment would make the agent react more quickly.

The BDI framework is closely linked to decision theory [105]. For example, the chance nodes in an influence diagram are a way to store the *belief*, composed of sensor data and the non-observable information that is inferred from them. The *desires* are embedded in the utility that is associated with a certain event. The world state that the agent wants to achieve, as well as its trade-offs between potentially conflicting goals, determine the value of each possible world state and, consequently, the utility of the available actions. Finally, *intentions* are determined at the decision nodes by maximizing the expected utility.

4.3.2 Communicative Goals

Cohen and Levesque [34] defined communicative actions for artificial agents in terms of the beliefs held by both parties. Because it is impossible to influence the interlocutor's mind directly, such actions represent *attempts* at changing their beliefs. An attempt, in turn, includes the *intention* to make an honest

effort at causing an *event* that leads to the desired change in the interaction partner's belief state.

For example, when a speaker requests a listener to do a particular action, the speaker's goal is to have the latter know that

- the speaker wants this action to happen in the future, and
- the speaker wants the listener to be the one doing this action.

Logical expressions can then be used to describe complex communicative intentions in terms of beliefs, action commitments, and temporal relationships.

In this thesis, however, communicative goals are modeled on a different level of abstraction. One reason is that the turn-taking model is supposed to be independent from the interaction domain and, consequently, from the type of information that the speaker provides. The intended change in the listener's knowledge - in other words, the function of the communicative act - does contribute to the urgency of a response, but the act's content plays a minor role. Therefore, a precise model of the participants' knowledge is not necessary.

Nevertheless, there is one core idea that this modeling approach from the artificial agent community shares with psychological literature about turn-taking. On the most basic level, the goal is to transmit information from one participant to the other. As detailed in section 3.4, it is possible that turn-taking signals started as the interlocutor's attempt at gaining or avoiding new information, which caused them to perform functionally necessary actions that have been ritualized over time.

4.4 Agent Frameworks

A number of software solutions exist for controlling ECAs. To avoid starting the implementation from scratch, several options were explored for building on existing work.

One thing that became apparent is a marked divide between the software used for virtual and robotic characters. While the application logic does not differ much between a graphics-based embodiment and one that is physically present, the associated technologies evolved in parallel, and different communities developed different solutions.

This sections summarizes the most relevant frameworks for interactive dialogue applications.

4.4.1 ROS

The Robot Operating System (ROS)[2] is an industry-standard and, therefore, a popular choice for both commercially available robots and research prototypes. Several platforms are built on this foundation (see table 4.1). Some researchers in social robotics have distributed their implementations as ROS nodes to increase the chances of them being reused [54, 57].

However, ROS is not readily available for all robots. Most of the platforms for which ROS drivers exist are functional rather than social, and the focus appears to be on practical capabilities such as navigation, computer vision, or object manipulation. Even in newer platforms that are built upon ROS, such as the Reeti V2 or the Aisoy KiK 1, the documentation focuses on the custom API wrapper instead of the ROS infrastructure.

Another problem is that ROS is hardly mentioned in the context of virtual agents, except when it comes to simulations of existing robot platforms. This means that, to develop turn-taking models for both virtual and robotic agents, an additional interface would be needed between ROS and the virtual agent.

developer	**platform**	languages	based on
Aldebaran	NAO 5	Python, C++, JavaScript	NAOqi (ROS available)
RoboKind	R-50	Java	GLUE.ai
	R-25	Java, C++	unknown (ROS available)
Robopec	Reeti V1	Java, C++	URBI
	Reeti V2	Java	ROS
Aisoy Robotics	Aisoy KiK 1	Python	ROS
Navel Robotics	Navel	Python	ROS

Table 4.1: Overview of the frameworks and APIs supported by the robots used at the University of Augsburg.

4.4.2 SAIBA

There have been concentrated efforts to unify multimodal behavior generation for graphical agents [72], and later robots as well [76]. The SAIBA framework ("**S**ituation, **A**gent, **I**ntention, **B**ehavior, **A**nimation") works in three distinct

[2] https://www.ros.org/

stages. First, the agent's *intent* is planned and represented in terms of communicative functions. It is transmitted to the behavior planning stage using the *Function Markup Language (FML)*. The selected behaviors and information regarding their temporal alignment are then sent to the behavior realization stage using the *Behavior Markup Language (BML)*. Kopp et al. [73] developed it further to work in an incremental context. The core features of their framework, which is called *Artificial Social Agent Platform* (ASAP), include incremental processing for both input and output, as well as a tight coupling between the planning and execution of commands. Frequent status updates keep the behavior planner informed about the information that has been delivered successfully, enabling a quick modification of the output in response to observed user behaviors or distractions in the environment.

Table 4.2 lists several SAIBA-compliant agent platforms. At the time of writing this thesis, all of them focus heavily on graphical agents. Le et al. [76, 75] did propose a way to extend the GRETA platform and use the same behavior generation for graphical agents and the NAO robot. Unfortunately, this extension is not found in the GRETA GitHub repository. Likewise, van Welbergen et al. proposed extending the ASAP architecture to social robots in 2014 [136]. The ASAP repository on GitHub indeed contains code for connecting to a NAO robot. However, they are labeled "experimental" and have not been updated since 2016.

Platform	**embodiment**		**last updated**
	graphical	robotic	
SmartBody [132] https://smartbody.ict.usc.edu/	yes	no	2017
BeAware [68, 69] no website given	yes	no	unknown
ASAP [73, 136] https://github.com/ArticulatedSocialAgentsPlatform/AsapRealizer	yes	experimental	2021
GRETA [100, 95, 76] https://github.com/isir/greta	yes	proposed	2022
Virtual Human Toolkit [51] https://vhtoolkit.ict.usc.edu/	yes	no	2022

Table 4.2: A list of SAIBA compliant agent platforms.

While there were at least experimental BML realizers for the NAO robot, no such realizers could be found for other robots such as the RoboKind R-50 or the Robopec Reeti. Nevertheless, the SAIBA framework provided a good starting point, and this thesis draws much inspiration from its principles.

4.5 Interaction Management Approaches

Many different approaches exist for implementing human-agent interaction, in particular dialogue systems. These include dialogue trees that select the next sentence based on a specified condition [109] or complex planners that choose dialogue strategies to fulfill certain goals. Those strategies could, for example, be derived logically from the agent's belief state [34] or based on rewards that were manually specified by the author [90].

This section focuses on two approaches that especially relevant for this thesis: finite state machines and conversational AI based on statistical language models. The prototypes presented in chapters 9 and 10 both used the former for modeling the dialogue that was then augmented by the turn-taking model. Additionally, the prototype in chapter 10 used elements of the latter approach to enable more flexible interaction with a human.

4.5.1 Finite State Machines

Finite state machines are a well-established approach for modeling agent behavior. It is also possible to nest state machines hierarchically or execute them in parallel, which allows for modeling highly complex behavior sequences.

Visual SceneMaker [46] is an authoring tool for such hierarchical and parallel finite state machines. It has been created by the German Research Center for Artificial Intelligence (Deutsches Forschungszentrum für Künstliche Intelligenz, DFKI). At the University of Augsburg, both virtual characters and an increasing number of different robots were controlled through this software around the time this thesis was started. This software was therefore chosen for the practical part of the thesis.

Figure 4.4 shows an excerpt of the state machine that was implemented for the interactive prototype in chapter 10.

States

A *state* represents a distinct step in the agent's behavior. Each one is either connected to a specific action, such as playing an animation or changing a variable's value, or it serves as a branching point in the program.

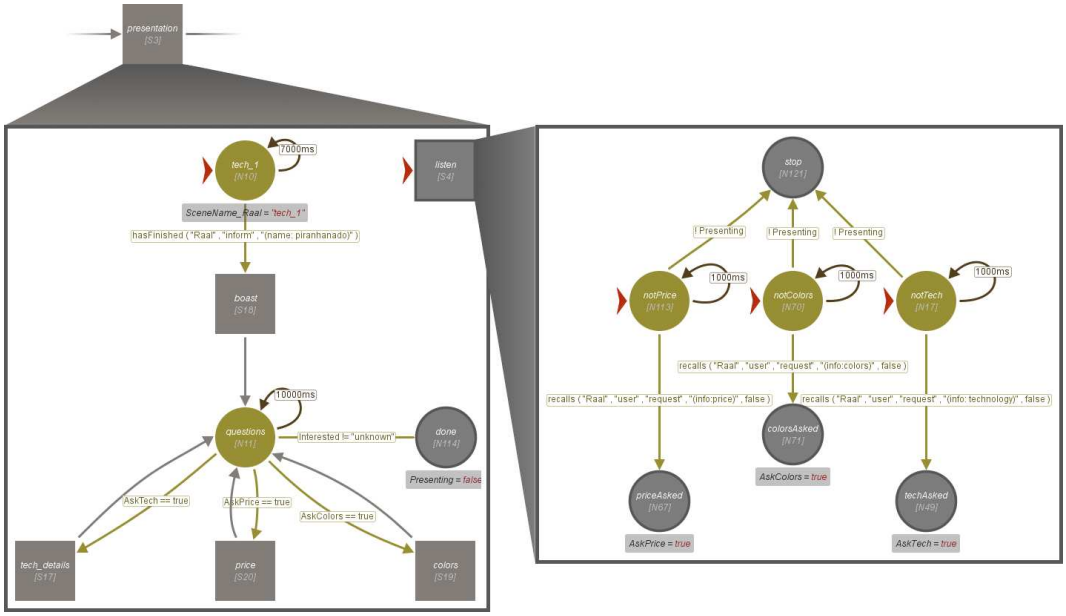


Figure 4.4: An example of a hierarchical finite state machine. The round nodes represent basic states whereas the square *super nodes* contain state machines of their own. The states are connected by *unconditional* (gray), *conditional* (yellow), or *timed* (brown) transitions. Red arrows mark the states that are active when the respective state machine is started.

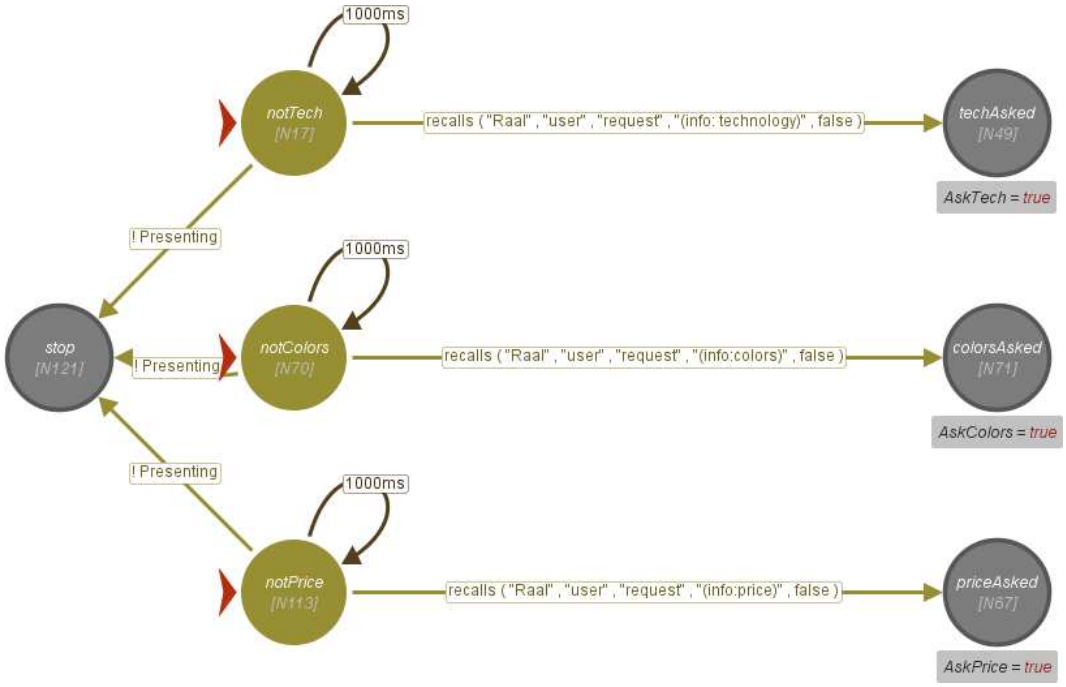


Figure 4.5: An example of a state machine with basic states. Nodes marked with a red arrow are activated in parallel when the state machine is started.

Figure 4.5 shows a state machine composed of basic state nodes. Three of them are marked as *start nodes*, which means that they are activated as soon as the state machine is executed. In this example, the start nodes serve as branching points in parallel processes. Each one has a process attached that monitors the conditions required for leaving this state.

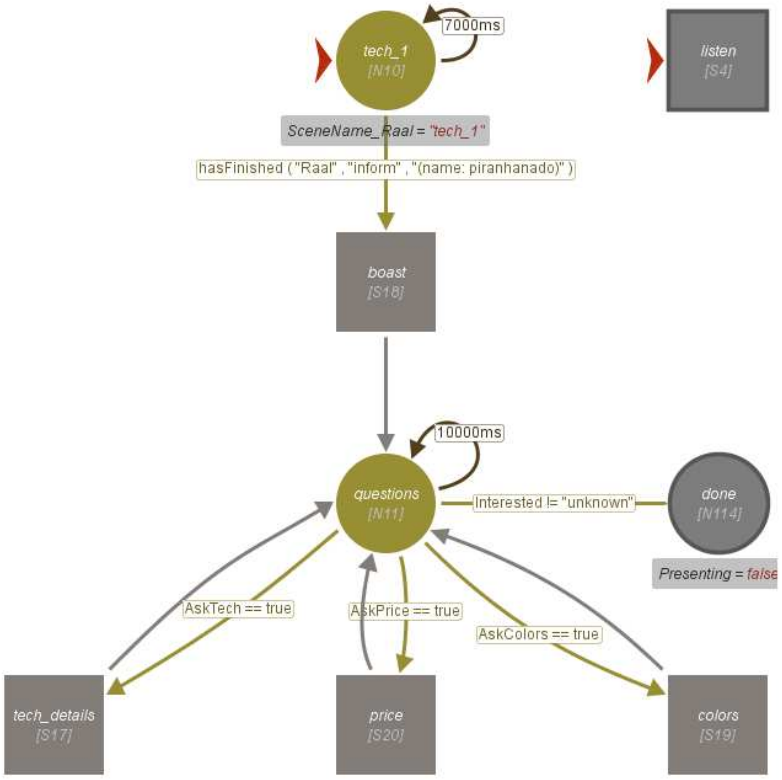


Figure 4.6: An example of a state machine that combines basic state nodes and super nodes.

Figure 4.6 shows a state machine that contains both basic nodes (round) and so-called super nodes (square). The latter contain other state machines and can be used for structuring the interaction hierarchically. For example, a conversation can be composed of phases such as exchanging greetings, dealing with different topics and saying goodbye. The subdialogue shown in figure 4.6 uses super nodes for going into detail about the current topic, the presentation of a vacuum cleaner that the implemented character is trying to sell. After introducing the topic at the start node, the next node holds a phase of boasting about the vacuum cleaner's technology. Questions from the interlocutor lead to phases that are dedicated to answering.

Transitions

A state machine is executed by moving from one state to the next, based on the conditions attached to the directed edges between the state nodes. Such conditions can be empty, in which case the edge is called an ϵ *transition*. Otherwise, the transition requires a given logical expression to be true. For example, the most recent speech input might need to provide a specific piece of information, or the time elapsed since the last input might have to exceed a certain threshold.

Visual SceneMaker offers different shortcuts for specific condition types. For example, figure 4.6 shows gray edges that represent ϵ transitions and yellow edges that are labeled with conditional statements. Some of these statements check the value of a given variable, whereas the edge leaving the start node calls a method in the attached program to perform a complex search in the semantic memory.

A third transition type, marked with brown edges, tests how many milliseconds have passed since the execution arrived at this node. The transition is triggered if this duration is longer than the one specified. Figure 4.6 shows how to use these timed transitions for waiting at a certain state until the conditions for proceeding are met.

Finally, Visual SceneMaker provides a special edge type for interruptive transitions. They lead out of a super node, interrupting the state machine within as soon as the attached condition is met. Those transitions are marked in red.

Actions

When a state is reached, any actions attached to it are executed. Such an action is typically a certain behavior that the agent will display, for example, an animation or speech command. Other actions may involve setting or calculating the values of certain variables that will be used in different states or branching conditions.

Visual SceneMaker, in particular uses its own scripting language and synchronization mechanisms for defining the agent's actions. Nonverbal behaviors, such as pointing gestures or head nods, can be directly inserted into the utterance at the point when they should be triggered. Since the source code of Visual SceneMaker is freely available, custom actions can easily be added. In this thesis, for example, all agents' semantic memories are updated as soon as the speaker has finished a certain part of its text.

Figure 4.7 shows an excerpt of the state machine used for the interactive prototype that is presented in chapter 10. To change the way speaking turns are scheduled, the action execution is decoupled from the dialogue flow using

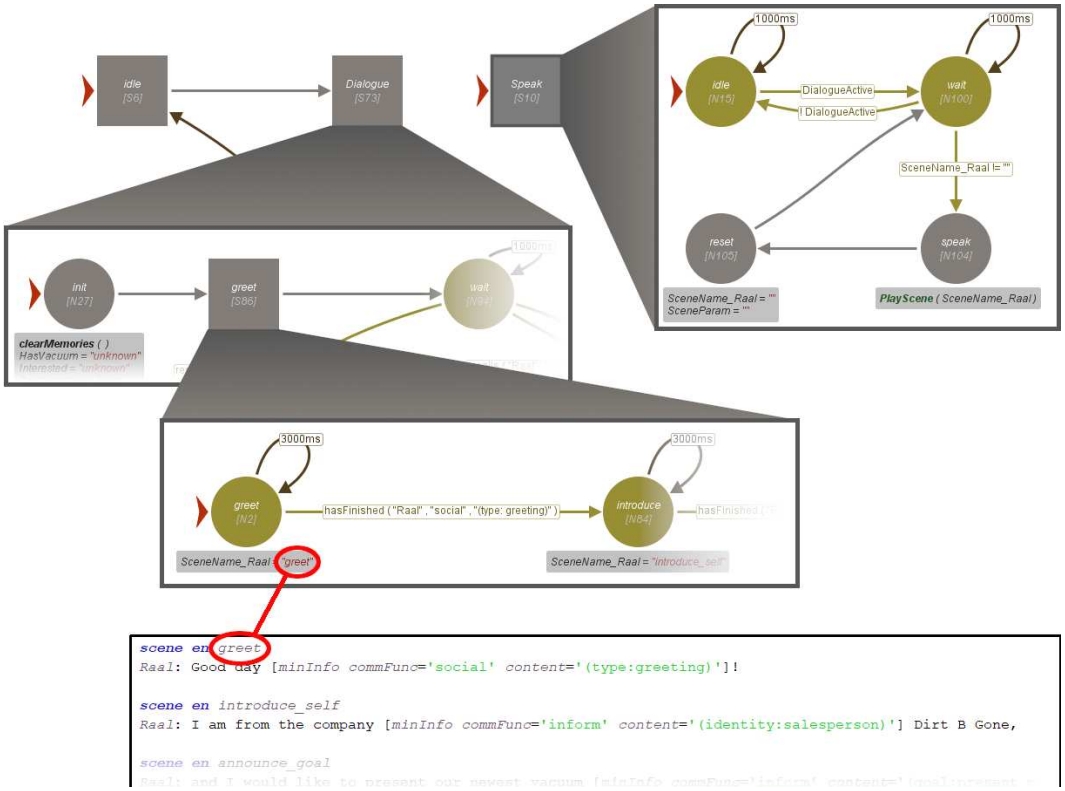


Figure 4.7: An excerpt of a hierarchical finite state machine that shows the actions attached to the states. Within the "Dialogue" state, a variable is set to the scene that the agent is supposed to execute. The parallel state machine "Speak" then uses this variable to execute this command asynchronously. Another action embedded in the scene itself then updates the agent's memory to let it know that the dialogue can advance.

a separate state machine that runs in parallel. States within the "Dialogue" machine set a global variable that specifies which scene the agent is supposed to execute. The parallel "Speak" machine waits for this variable to be set and then triggers that execution. The scene itself is defined in a separate script and has a secondary action embedded in the text. When the agent speaks this text and reaches the embedded action marker, its content is used to update the agents' memory in the background. Consequently, the process executing the "Dialogue" machine can successfully retrieve that memory and proceed to the next state.

4.5.2 Conversational AI

In recent years, there has been a growing interest in using machine learning to make conversational agents more flexible and intuitive to use.

For example, Google offers the Dialogflow[3] cloud service for designing, training, and deploying dialogue systems that are based on artificial intelligence. A popular alternative is Rasa[4], a software framework that is available via paid services or as an open source version[5].

Input Processing

Both Google Dialogflow and Rasa need a set of example phrases that are associated with a so-called *intent* and optional *entities*. These terms are roughly equivalent to a dialogue act's *communicative function* and its *content* in the DiAML standard (see section 2.4.2).

Google recommends about 10 to 20 example phrases per intent on order to prepare the conversational agent for a sufficiently wide range of user inputs[6]. Both frameworks use pre-trained language models to match similar input sentences to the examples, which leads to a greater tolerance for grammatically incorrect or misspelled inputs in comparison to using a rule-based grammar.

Conversation Structure

A "flow" in Google Dialogflow "can be described and visualized as a state machine" according to their official documentation[7]. As with hierarchical state machines, the conversation can be structured into different interconnected phases (called "pages" in Google Dialogflow) that provide and collect specific information. It transitions to different pages based on the user's recognized intent or on events such as recognition failures.

For Rasa, the interaction designer provides so-called "stories"[8] that consist of example sequences of system actions and user inputs. It is also possible to define more rigid rules for specific sub-dialogues, but the documentation advises using that option sparingly in order to make the agent more robust.

Both frameworks also have the concept of "forms"[9,10] that serve to gather a set of parameters for fulfilling the user's request. For these forms, the interaction designer specifies a list of parameters that need to be collected along with the utterances that the agent will use to ask for each of them.

After those example conversations are defined, both frameworks use them to train the agent's dialogue management model. Machine learning is used

[3] https://cloud.google.com/dialogflow
[4] https://www.rasa.com/
[5] https://rasa.community/
[6] https://cloud.google.com/dialogflow/cx/docs/concept/intent#tp
[7] https://cloud.google.com/dialogflow/cx/docs/basics#page
[8] https://rasa.com/docs/rasa/stories
[9] https://cloud.google.com/dialogflow/cx/docs/concept/parameter#form
[10] https://rasa.com/docs/rasa/forms

to generalize from the provided interaction sequences to similar but unknown sequences of the participants' utterances.

Agent Actions

Regarding the agent's actions, there are many similarities between conversational AI solutions and the finite state machine approach.

In all of them, the steps of the conversations are linked to phrases that the agent should speak or to actions such as setting variables or retrieving information from a database. In both Google Dialogflow and Rasa, as well as in Visual SceneMaker, the interaction designer can provide alternative phrases for the same action so that the agent's speech becomes less repetitive.

Specifying the agent's utterances explicitly has the advantage of giving the designer full control over its responses, albeit at the cost of flexibility. However, it should be noted that recent years saw the rise of generative AI technologies, such as ChatGPT[11] that uses a so-called *large language model (LLM)* trained on massive datasets of human-authored content. These complex models allow for creating or rephrasing text with far less effort on the developer's side. For example, Axelsson and Skantze [12] used the LLM GPT-3 to automatically translate raw facts from their knowledge base to natural-sounding sentences and remove redundancies such as repeated references to the same subject.

Google advertises a separate service called "Vertex AI"[12] that automatically analyzes a given website and creates a Dialogflow agent capable of having a conversation about it. Rasa's documentation indicates that it can be integrated with third-party LLMs, although this integration still appears to be experimental at the time of writing this thesis[13].

4.6 Agent Implementations

So far, this chapter has looked at the approaches for coordinating one or more agents' behavior between each other or with human users. However, it is also important to look at the way individual agents are implemented and how they can be controlled by the interaction management software.

Both the hardware and software of artificial agents are highly heterogeneous. Capabilities and the degrees to which they are made accessible vary between manufacturers and sometimes also between different robot models or software versions.

[11] https://openai.com/chatgpt
[12] https://cloud.google.com/generative-ai-app-builder
[13] https://rasa.com/docs/rasa/next/llms/llm-nlg

This section presents the most important differences and the challenges they pose for implementing a behavior model that can be used with different agents.

4.6.1 Programming Languages

One major challenge for re-using code with different agents was that they are programmed in different languages. As shown in table 4.1, about half the robots come with a Java API, whereas Python is the language of choice for the others. In those cases when different languages are supported, they do not necessarily offer the same functionality, or they may come with additional overhead for starting the software.

For example, when the Reeti V1 was first released, the C++ API was recommended by the manufacturer because it was more thoroughly tested than the Java API. Extension modules written in C++ could also be loaded directly into the Urbi environment on which the robot was built, whereas those written in Java required a more complicated workaround.

As for graphical agents, the ones available during this thesis were programmed in game engines that used either C++ (such as the Horde3D engine[14] that was developed at the University of Augsburg) or C# (such as the freely available Unity engine[15]). Charamel's VuppetMaster agent is loaded into HTML pages and controlled via JavaScript[16].

4.6.2 Labels, Units and Values

Even those agents that support the same programming language cannot directly be controlled by the code developed for another. Different development teams have decided on different naming schemes, units, and value ranges for their respective agent, which makes it necessary to convert certain command parameters when connecting the dialogue manager to a different platform.

For example, a critical prerequisite for turn-taking are the bookmarks that have to be inserted at the end of the minimum necessary information (MNI) (see section 5.2.1). Depending on the text-to-speech (TTS) software that a particular agent used, they must be provided in a different syntax. A comparison of these bookmark formats can be found in table 4.3.

When it comes to moving specific joints, for example, to animate the agent's gaze, almost every agent platform is different. Graphical agents commonly take a set of coordinates and use inverse kinematics to derive the appropriate

[14] http://horde3d.org/
[15] https://unity.com/
[16] https://vuppetmaster.de/documentation/docs/overview/gettingstarted

developer	**platform**	TTS Engine	bookmark format
Aldebaran	NAO 5	Acapela	*mrk* = 42\
RoboKind	R-50	Acapela	*mrk* = 42\
	R-25	Acapela	*mrk* = 42\
Robopec	Reeti V1	Loquendo	*book* = 42
	Reeti V2	Loquendo	*book* = 42
Aisoy Robotics	Aisoy KiK 1	Festival, eSpeak, Pico	-
Navel Robotics	Navel	Acapela	*mrk* = 42\
Charamel	VuppetMaster	Amazon Polly	-

Table 4.3: TTS software used by different agent platforms.

join rotations automatically. Social robot APIs rarely provide this option. Of the robots used at the University of Augsburg, only the Navel robot uses this animation approach at the time of writing.

Furthermore, few social robots follow the joint naming conventions of their industrial counterparts. Of those used in the context of this thesis, only the NAO and RoboKind robots use the "pitch/roll/yaw" convention. However, the joint names still differ in spelling, since the NAO uses camel case and the RoboKind R-50 uses all capital letters with underscores. Table 4.4 shows a subset of these differences.

developer	**platform**	Neck Joints	Position	Velocity
Aldebaran	NAO 5	HeadPitch, HeadYaw	angle [min; max] °	speed [0.0; 1.0]
RoboKind	R-50	NECK_PITCH, NECK_YAW	linear [0.0; 1.0]	time [0; ∞[ms
Robopec	Reeti V1	neckTilt, neckRotat	linear [0.0; 1.0]	time [0.0; ∞[s
	Reeti V2	neckTilt, neckRotat	linear [0.0; 100.0]	speed [0; 100] %
Aisoy Robotics	Aisoy KiK 1	head_v, head_h	linear [0; 100]	time [0.0; ∞[s

Table 4.4: Neck joint names, units and value ranges used for animating different robot platforms.

That table also compares the vastly different parameter values for animating these joints. Only the NAO can be controlled by specifying the joint angle, whereas the others require a position in a linear range. However, the scale for this position is not consistent, either. For the Robopec Reeti, it even differs between consecutive models.

Similar issues affect the animation duration. Some APIs allow the developer to specify the exact time after which the animation should be finished, either in seconds or milliseconds. Others, however, only let them select the desired fraction of the joint's maximum speed.

These inconsistencies between different platforms are further complicated by a lack of documentation that will be detailed in section 4.6.3.

4.6.3 Software Limitations

Some functionality is limited by the available hardware, such as the number of degrees of freedom (DOF) in a robot's neck or the processing power for generating audio from a given text. However, most limitations appear to come from the software that is used to control it.

Lack of Control

Access to the agent's output capabilities is limited by the available API, and many are controlled by proprietary software that does not necessarily expose all features. Consequently, developers often depend on the manufacturer to make the necessary functionality accessible.

For example, the RoboKind R-50 software does not support cancellation of speech jobs that have already been started, although this functionality exists for the Aldebaran NAO that uses the same TTS software. At the time of writing this, the Aisoy KiK 1 executes speech and animation tasks in a blocking manner, although the movement of the neck servos should logically be independent of the audio output. As for graphical agents, the VuppetMaster platform by Charamel only provides pre-defined animation files, but allows no direct manipulation of the virtual character's skeleton[17].

Lack of Transparency

Similar problems exist when it comes to monitoring the execution progress. Not all platforms provide detailed callback mechanisms, feedback events, or easily queried status variables. This can make it necessary to implement additional scheduling mechanisms, such as explicitly blocking the execution of incoming commands for the expected time of the current one.

[17] https://vuppetmaster.de/documentation/docs/api/commands

For example, the RoboKind R50 robot does not expose the current status of an animation job, and early versions of the Reeti V1 API did not forward the bookmark events of the Loquendo TTS software. Likewise, the virtual characters on Charamel's VuppetMaster platform do not offer bookmark events, although the Amazon Polly TTS engine would support them[18].

Lack of Documentation

Even the best software is useless if nobody understands how to use it. Unfortunately, many agent platforms are insufficiently documented. Newly developed social robots, especially from not-yet-established companies, are especially prone to suffering from this. Similar issues are likely with research prototypes that are mainly used by their original developers.

There is often a lack of information on available classes, methods and their parameters. For example, it is not always evident whether a servo position needs to be given as an angle, a fraction of 1.0, or a percentage. Some animation components return handles to the animation job, but others only provide events that need to be subscribed to in a specific way. Also, the documentation rarely mentions what will happen when two contradictory commands are given simultaneously and whether there is already a form of conflict handling implemented.

Most robots are animated by moving a servo to a fraction of its range without associating that range with clearly defined angles. Unfortunately, the robots' documentation rarely states the angles to which these ranges are mapped. Of those robots used at the University of Augsburg, only the NAO comes with detailed information on angles and joint arrangement. For the RoboKind R-50, a similar joint map could be requested from the manufacturer, but it did not match the actual model that was in use at the chair. Consequently, most of the robots' angle limits had to be measured by hand before they could be used for procedural animation.

4.7 Conclusion

This chapter gave an overview of the technical background for this thesis.

A decision-theoretic approach was chosen because it represents an idealized version of human decision-making. In particular, the concept of multiattribute utilities will play an important role in the presented turn-taking model. It allows for more systematic reasoning about the consequences of the chosen behaviors by breaking the outcome down into concrete factors, such as the

[18] https://docs.aws.amazon.com/polly/latest/dg/supportedtags.html

obtainable information and the importance that this information has for a given personality. Therefore, the mathematical principles behind this approach were summed up briefly. For more detailed explanations, it is recommended to look at "Learning Bayesian Networks" by Neapolitan [94] and "Foundations of Multiattribute Utility" by Abbas [1].

Section 4.3 looked at the systematic representation of semantic knowledge in the interaction between humans, conversational agents, or any combination thereof before section 4.4 summarized the two most relevant frameworks for computer-controlled agents. This thesis will build upon both the DiAML standard that was established for annotating communicative acts [22, 101] and the BML representation of agent actions in the SAIBA framework [72]. Another major aspect that will be adopted from SAIBA is the clear separation between the agent's intention and its surface behavior. Its extension, the ASAP framework [73], provides the principle of continuously monitoring the execution progress of these behaviors. Chapters 7 and 8 will explain in more detail how this thesis builds on what was reviewed here.

Two current approaches for implementing human-agent interaction were summarized in section 4.5. While finite state machines ended up playing a greater role in implementing the proof-of-concept applications (see chapters 9 and 10), there is also an overview of conversational AI frameworks that have recently grown in popularity. As will be shown in section 10.3.4, the final prototype of this thesis uses Rasa's NLU component in order to provide incremental speech input.

Finally, this chapter examined the commonalities and differences of currently available agents, both graphically embodied and robotic. It pointed out some technical challenges that must be addressed when building a generalizable, agent-agnostic turn-taking model. In this thesis, the RobotEngine framework (see chapter 8) was developed for this very purpose.

The next chapter will have a closer look at existing research in the field of ECAs, their interaction with humans, and the factors that shape humans' perception of artificial characters.

Chapter 5

Related Work

5.1 Introduction

Recent years have seen much technological progress regarding the decoding and encoding of speech and nonverbal signals. Many behavior patterns observed in human communication have already been transferred to the interaction with artificial characters. Experiments often show that people respond to the behaviors of a virtual or robotic agent as they would to those of a fellow human. On those occasions when the artificial behaviors fail to achieve the desired result, such experiments can reveal gaps in the underlying theories, possibly leading to a better understanding of human communication as well.

While consumer products are still limited to a very rigid turn structure, various research institutes have been working on making human-agent interaction more life-like. For example, incremental processing lets agents react already during the user's input. Other research has focused on dynamic, situation-appropriate behavior generation and expressing different personalities through variations in said behavior.

These topics may appear very disconnected at first glance, but they all contribute to the same goal of implementing fluent and intuitive communication between humans and ECAs. Research on incremental systems focuses on the question of *when* the agent should respond, whereas research on behavior answers the question of *how* the agent responds. Personality simulation examines the *effect* that this response has on the observer's opinion of the agent.

This chapter will review existing research on turn-taking and personality simulation in human-agent interaction. First, it will look at approaches for determining the best moment to respond. The next section will then cover nonverbal behaviors that have been linked to internal states, both for understanding the user's intentions and expressing those of the robot. Finally, there

will be a section about adaptable and adaptive conversational agents because finding a general solution for a large user base is challenging.

5.2 Action Timing

Turn-taking conflicts happen when a response comes at an inappropriate time, for example, before the current speaker has finished their turn or after an awkward pause. Therefore, several works have examined when an agent should respond to the user. Some also tried to predict when the user was likely to respond to the agent or to distinguish between actual barge-ins and backchannel signals. This information enabled the agent to decide between yielding or holding the turn.

5.2.1 Meaningful Prefix

Most conversational agents, especially those in consumer products, only respond after the user's input has been completed. Incremental dialogue systems aim to solve this problem by recognizing the user's intention before that point. They revise their hypothesis about the user's intent with every new piece of information, such as the next spoken syllable or gesture stroke. Consequently, several related works cover the issue of detecting the meaningful prefix of an input that has to be processed before responding even becomes an option.

DeVault, Sagae, and Traum [38] trained a decision tree model on a domain-specific corpus of natural language utterances to detect the so-called "points of maximum understanding." They reasoned that their virtual character should not try to complete a sentence if there was a chance that it could understand the user better by listening for a longer time. The features on which they trained their classifier were based on, for example, the length of the partial input, the probability distribution for all possible final results, and the probability of the most likely input. At runtime, the classifier then determined the point at which the agent had heard enough to respond. To complete the user's utterance, the partial input was mapped to the sentence in the corpus that had the most similar prefix, and the agent then spoke the remainder of the sentence.

In a later implementation of a similar application, Traum et al. [135] likewise used a statistical classifier to predict two semantic frames, one filled by the expected full input and one by the already recognized prefix. Visser et al. [137] then used the confidence in the classifier's result for deciding when the agent should respond by, for example, frowning in confusion, nodding attentively, or completing the partial sentence.

Chao [26] defined the concept of *minimum necessary information* (MNI for short) as "the minimum amount of information needed to be conveyed by the robot for the human to respond in a semantically appropriate way." The data she analyzed for human-robot dyads showed that the end of the MNI reliably predicted the earliest point at which a human would start to respond. She concluded that the communicative goal was achieved as soon as the MNI had been successfully transmitted and common ground had been established (see section 3.4). Consequently, the robot could save time by stopping after that point, making the conversation more fluent and efficient. In contrast, being interrupted before the MNI would require the robot to try communicating this information again. However, Chao noted that determining the end of the MNI was not trivial. In one of her example scenarios, a game of "Simon says" with low complexity, this information was manually associated with possible actions that either the human or the robot could perform.

Skantze et al. [122] examined several turn-taking cues that could be employed by a robot providing instructions to a human. One of those cues was the lack of syntactic completeness. Their experiment showed that, at least when the participants could see the robot's face, they were more hesitant to respond to syntactically incomplete utterances. They were less likely to give verbal feedback, more likely to look at the robot during the following pause, and slowed down the drawing activity through which the robot was guiding them.

This thesis adopts Chao's term "minimum necessary information" for the initial part of a message after which an interruption is likely to occur [26].

5.2.2 Response Timing

Even when there is enough information to act on, there is still the question of when exactly the response should follow. As explained in section 3.4.3, interruptions and overlaps carry information regarding the social relationship of the speakers and can come across as impolite in certain contexts. Gaze also plays a major role in coordinating speech because it indicates when the other party is ready to listen. For this reason, several works incorporate gaze information into the decision process, while others also consider the interlocutor's affective state.

Resource Management

Chao [27, 26] defined turn management as a resource management problem in parallel processes. Her approach was to model it with Timed Petri Nets to enforce that several conditions were met before a response. For example,

to gaze at a given target, the agent needs to use the same joints as it does for a nodding animation. To speak, it needs to hold the conversational floor while the user does not. To manipulate objects in a shared workspace, the target object must not be in the hands of the user. This approach allowed not only for coordinating the agent's turns with those of its partner but also for synchronizing the agent's own behaviors across modalities.

When two participants compete for the same resource, this creates a turn conflict that needs to be resolved via different means. In Chao's work, this was done by maintaining a pre-defined ratio of speaking time between the robot and the human, as well as by a set of configurable time thresholds. One experiment that compared different threshold configurations also made the *passive* robot wait for the human to make eye contact before the robot was allowed to take the turn. In other words, taking the turn required both an unoccupied voice channel and the visual attention of the interlocutor.

Timeouts

Duration thresholds have long been used to detect opportunities for the agent's response and to model engagement during human-agent interaction.

Rich et al. [108] defined several so-called *connection events* after observing collaboration between humans. They then implemented the automatic recognition of such connection events for monitoring the engagement in human-robot collaboration, and in a later work [54], they used the defined mechanisms to generate such events when the engagement was low. Events were only considered successful if the required response followed within a given time window. The thresholds for these delays were subjectively determined by the authors.

- **Directed Gaze:** One participant turns their visual attention to an object in the shared workspace. The other follows their line of gaze to look at the same object. The maximum delay was set to 3.0 seconds.
- **Mutual Facial Gaze:** One participant looks at the other's face, and the latter responds in kind. The maximum delay was set to 1.8 seconds.
- **Adjacency Pair:** One participant contributes to the interaction by, for example, making a statement, asking a question, or performing an action. The other gives a response that continues the interaction, such as a comment on the statement, an answer to the question, or the next step of the shared task. The maximum delay was set to 3.1 seconds.
- **Backchannel:** One participant reacts briefly while the other performs a longer action or utterance. No delay was involved in this case.

A different pattern, the completion of the speaker's sentence, was implemented by DeVault et al. for a negotiation training simulation [38]. Such completions can serve as a grounding mechanism (see section 3.4) by showing that the listener is paying attention to the topic and can infer what the speaker intends to tell them. The virtual characters in that simulation had been given the ability to predict the remainder of the human's utterance based on a statistical model trained on example sentences. To avoid offending the user by barging in, the agents were programmed to wait for 600 milliseconds of silence before completing the phrase with the predicted text.

Visser et al. later applied this pattern of utterance completion to a different training scenario [137]. In that setup, the agents had been equipped with more complex rules for responding to the user. First, the agent determined whether the human was speaking and nodded if the NLU component had detected input with high confidence. However, if the human was silent, the agent's reaction depended on the duration of the pause. Pauses of about 200 milliseconds led the agents to signal their level of understanding, ranging from a confused frown to nodding and, in case of full confidence in the glsnlu result, a verbal backchannel. If the pause was longer than 600 milliseconds, the agent reacted verbally by either completing a partial utterance or responding to a complete one.

None of these works explicitly modeled the personality of the agent. However, Chao [26] used variations in the timing threshold to have the robot show different levels of Extraversion and attentiveness.

- **Lapse Tolerance:** The *active* robot configuration tolerated pauses of up to 500 milliseconds between turns, whereas the *passive* one waited for up to 4000 milliseconds.
- **Act Spacing:** Within a turn, the robot produced up to three segments of gibberish speech. The *active* one left gaps of 50-250 milliseconds between these, whereas the *passive* one left gaps of 200-1000 milliseconds.
- **Backchannel Spacing:** The gap between vocal backchannels was 2000-4000 milliseconds for the *active* condition and 4000-6000 milliseconds for the *passive* one.

A user study was then conducted in an interactive setup, comparing the personality that humans associated with the robot in the two conditions. The *active* robot configuration was rated as significantly more extraverted than the *passive* one. Furthermore, the study participants were allowed to label their personalities freely. For the *active* robot, these labels included "outgoing/extroverted" and "bold/confident", but also "aloof/distant" which could

hint at a more self-centered personality impression. The *passive* robot was labeled with words such as "shy" and "unresponsive/silent", reflecting the low Extraversion rating. The labels also included "moody/temperamental", indicating a difference in perceived Neuroticism between both configurations. The latter difference supports the idea that Neuroticism may influence observable behaviors (see section 3.2.4).

Threlkeld, Umair, and de Ruiter [133] fitted Bayesian models to the periods of silence in an audio-only corpus. They obtained probability distributions for which silence duration precedes a speaker switch and which one precedes continuation by the same speaker. Based on these distributions, they calculated the probability of a speaker switch given the elapsed time since the end of the voice activity. Furthermore, they identified several timeouts for these switches.

- **Start the response:** The probability of a speaker switch is highest after about 150-200 ms of silence. The response should be started before 394 ms because both sides are equally likely to start speaking at that point.
- **Prompt for a response:** If the gap is longer than 762 ms, the other participant is least likely to take the yielded floor, so the last speaker should explicitly ask them to respond after that point.

Decision-Theoretic Approaches

Timeouts are not always appropriate in practice. First, it is hard to determine fixed duration thresholds that apply equally well to different situations. Second, as Skantze pointed out in a 2021 review on turn-taking [121], *pauses* within a turn are often longer than the *gaps* which occur when the speaker changes. To mitigate those problems, some researchers used decision-theoretic approaches for determining the best response time based on context information.

Horvitz, Koch, and Apacible [56] implemented a system called "BusyBody" which could be trained to predict the cost of interrupting the user. During training mode, users were repeatedly prompted to state whether they were busy at this moment or not. The answers were then combined with activity logged on the computer and conversations detected in the user's room. Finally, a Bayesian network was trained from the co-occurrence of states to predict when the user is busy and would be bothered by a notification.

A later dialogue system by Bohus and Horvitz [18, 17], taking the form of a virtual moderator at a quiz kiosk, relied on utility-based reasoning for calculating the waiting time before taking the turn. Interactive systems need to account for various uncertainties, such as ambiguous inputs or non-deterministic

delays in the processing pipeline. Therefore, the authors set up a detailed probabilistic model to determine who had been addressed by the last speaker, how long the system's reaction time would be, and how likely it was that one of the humans in front of the kiosk would step in during that reaction delay. Additionally, they had human raters assign costs to different types of turn-taking errors, such as taking the turn when it had been yielded to another human and thus speaking at the same time as the assigned speaker. These costs were then used to calculate the expected utility of various waiting times and to choose the time that minimized the cost.

On a coarser time scale, Conati [35] proposed that a virtual butler should consider the uncertainty in assessing the user's needs and take both the costs and benefits of each action into account. Specifically, she described the use of a *dynamic decision network* for inferring the user's affective state and possible causes for it from the observable surface behaviors. Emotions were to be inferred from the user's goals via the OCC model [98], to ensure that the system could distinguish clearly between emotions that were very similar in outward expression. For example, if the user experienced shame, the agent would need to bolster the user's confidence, whereas it would need to apologize if the user experienced reproach. Furthermore, by predicting how certain messages would impact the user's affective state, the agent would be able to delay those that would be harmful in the current circumstances.

Conati also suggested that the relevant goals could depend on the user's personality traits. Said traits would then influence interaction patterns and observable emotional expressions via those goals. Another advantage of such a model would be that the system could be used in both a *predictive* and a *diagnostic* manner. In other words, knowledge about the user's personality traits and their observable behavior could both be used to reason about their goals. Finally, Conati pointed out that the decision-theoretic model could increase the transparency of the agent's timing decisions. Thanks to the explicit causal relationships, the system would be able to explain its decisions, making the agent more trustworthy and its actions more acceptable.

5.2.3 Explicit Signals

Rich et al. [108] identified several behavior patterns related to engagement in a collaborative task. They modeled their so-called "connection events" in human-robot interaction with a combination of duration thresholds and explicit signals, specifically gaze shifts, pointing gestures, and speech activity. One participant - either the human or the robot - initiated the connection event and waited for the other to respond appropriately. The connection event failed if the expected response did not follow within a specific time window.

- **Directed gaze:** When one participant points or looks at an object, a "directed gaze" event is recognized after the other participant looks at the same object.
- **Mutual facial gaze:** When one participant looks at the other's face, a "mutual facial gaze" event is recognized after the latter participant looks at the first one's face as well.
- **Adjacency Pair:** When one participant yields the floor after speaking, an "adjacency pair" event is recognized after the other participant takes the floor.
- **Backchannel:** While one participant has the floor, a "backchannel" event is recognized after the other participant nods, shakes their head, or speaks.

A later work by Holroyd at al. [54] implemented a behavior generation module based on those connection events. Several policies defined at which point of the interaction the robot would initiate or respond to the respective events. For example, the robot ended its speaking turn by looking at the human, followed their gaze to objects in the workspace, and made eye contact when the user looked at it. The main purpose of this module was to make the robot appear engaged in the shared task and encourage the user to be engaged as well.

Own prior work with colleagues [83] involved modeling a typical gaze pattern that is associated with handing over the conversational floor. In this setup, a human interacted with a social robot via speech, gaze, and placing objects on an interactive table. After a question was recognized in the verbal channel, the robot waited until the user's head had turned towards it. It then made eye contact before answering. A timeout rule (as described earlier) ensured that the robot would answer eventually, even without detecting the expected gaze shift.

5.2.4 Learned Response Behavior

Chýlek et al. [30] trained and compared several classifiers for predicting the appropriate time for interrupting the speaker, based on the interlocutors' prosody. They used two corpora of unrestricted conversations in which the voices of both speakers were recorded on separate channels. Both corpora had been manually annotated with time stamps of overlapping speech. To generate negative examples for training, the authors reasoned that the interrupting person had intentionally avoided speaking in the time window preceding

the overlap. Their results showed that the accuracy could be improved by not only considering the presence of an overlap but also whether it led to a speaker change. However, their classifiers only reached accuracies between 61 and 69% percent. The authors noted that although this accuracy was better than chance, it would not be sufficient for practical use.

Withanage Don et al. [139] used machine learning to extract listening behavior patterns from a corpus of therapist behavior. The trained model was used to predict the agent's behavior in real-time based on the user's voice activity, prosody, facial expression, and gaze direction during the previous 1000 ms. The generated behavior encompassed facial expressions as well as the direction of the head and eyes in a standardized form that could be displayed on different graphical agents. While timing is not mentioned explicitly, the authors describe a system that can alter the agent's behavior with a delay of approximately one second in direct response to that of the user.

Taillandier et al. [128] had pairs of agents develop turn-taking strategies from the ground up. The agents were supposed to solve a task together, while communication was limited to a single message channel. Furthermore, different approaches were tested for simulating the lack of clarity caused by overlapping speech. In case of overlaps, the message from the partner was either converted to random noise, substituted with a different message, or canceled out entirely. The agents were implemented as single layer long short-term memory (LSTM) networks and trained to optimize their policies via reinforcement learning. Results showed that, while not all agent pairings were able to develop appropriate turn-taking strategies, those who did solved the shared task with higher accuracy. However, to agree on a turn-taking strategy, the agents must be aware that overlaps are occurring. Consequently, agents that received altered messages performed better than those whose speech canceled out that of the partner.

Yang, Achard, and Pelachaud [140] trained an support vector machine (SVM) on the multimodal NoXi corpus [25] to predict interruption timing. They used several modalities that had been automatically extracted with the help of state-of-the-art third-party classifiers. Specifically, they extracted the volume and prosody from the audio, as well as the facial expressions, gaze direction, and head movements from the video. Using the approach of Chýlek et al. [30], they also trained two different classifiers on the NoXi corpus and found that their multimodal SVM classifier performed best. They then conducted a perception study in which a static image of two cartoon characters was combined with overlapping audio. Study results showed that random interruption timing was perceived as very similar to predicted and ground truth timing, and significant differences only emerged regarding other variables.

- **Interruption type:** Agreement interruptions were perceived as better placed and more acceptable than disagreements. Interrupters were perceived as more cooperative, more friendly, and less competitive than disagreeing ones. However, agreements and clarification requests made the interrupter appear more likely to control the conversation and grab the turn.

- **Voice type:** Interruptions were perceived as more acceptable, less competitive, and less dominant when a human voice recording was used than in the case of synthetic TTS audio. Synthetic voices were more strongly perceived as interrupting unnecessarily.

The results of Yang et al. [140] confirm that humans judge artificial agents more strictly than humans (see also [78]). They also show that audio timing alone matters less than the context in which an interruption occurs. It should further be noted that their prediction of interruption timing was based on both audio and video data, whereas their perception study did not provide participants with any visually observable behaviors. This loss of information might explain why the predicted timings did not perform significantly better than randomly-timed interruptions.

5.3 Behavior Reflecting Internal States

Numerous researchers have successfully transferred human behavior patterns to artificial agents to improve the interaction. Their results often show that people interpret those patterns as if they were displayed by another person, using them to intuitively deduce what is happening in the agent's "mind". This information helps establish the common ground regarding both the process and the content of their communication (see section 3.3.1).

This section will detail how nonverbal behaviors have been used to signal attention, turn-taking intentions, and affective states.

5.3.1 Attention

Several studies have shown that turning a robot's eyes toward persons or objects of interest makes it easier for humans to understand its intentions. Such gaze patterns not only serve to resolve referential ambiguities but also indicate whether the robot is interested in the interaction or deep in thought.

Referential Target

Staudte and Crocker [127] showed that people are more likely to fixate on the same object that the robot appears to look at. Furthermore, they were quicker to determine whether a statement was true when the robot looked at the same objects that it was talking about. In a similar experiment, Häring, Eichberg, and André [59] showed that referential gaze can improve the human's performance in a collaborative task. The task in question was an abstract puzzle game to which only the robot knew the solution. When the appearance of the puzzle pieces was ambiguous, having the robot look at the referenced object and the target location led to the participants making fewer errors.

Huang and Thomaz [58] examined whether the robot's gaze influenced the performance during a teaching task. Study participants interacted with a humanoid robot, using speech and pointing gestures to label building bricks with a color and a name. When the robot followed the pointing gesture with its gaze or looked at verbally referenced objects, the participants were significantly more efficient in teaching it. Fewer misunderstandings occurred, and fewer steps were necessary to resolve them. Furthermore, they considered the robot a better collaboration partner than when it only looked at the participant. They found it easier to determine if the robot had understood them, which was also reflected in the reduced number of redundant labels and confirmation questions. Finally, the robot was perceived to be more intelligent, life-like, and engaged in the task.

Skantze et al. [122] showed that study participants were faster at drawing a path on a map when the Furhat robot gazed at the landmarks it was referencing in its verbal instructions. Compared to a condition with random gaze behavior, ambiguous references were resolved faster, fewer misunderstandings occurred, and the participants rated the gaze as helpful rather than confusing.

Engagement

As mentioned before, Rich et al. [108] identified several patterns of speech and gaze that are important for engagement in conversation. They implemented a module for detecting directed gaze, mutual facial gaze, adjacency pairs, and backchannel signals in human-robot interaction. The same authors [54] later used these "connection events" to monitor the level of engagement and repair it when necessary. If more than the average amount of time had passed since the detection of such an event, the robot's policy was to glance at the human's face briefly.

During face-to-face interaction, humans constantly monitor the other person to ensure their continued attention to themselves and objects of interest. Therefore, Huang and Thomaz [58] tested whether a robot was seen as a bet-

ter communicator when it did the same. People were shown video clips of the robot presenting information, delivering a message, or giving directions. Overall, they preferred the videos in which the robot looked back and forth to ensure that the interaction partner was paying attention and paused the interaction when the human was momentarily distracted.

Baur et al. [16] used a dynamic Bayesian network to infer the user's engagement during human-agent interaction. They modeled both the *individual engagement* that was expressed in the user's body language and the *interpersonal engagement* that was associated with the connection events defined by Rich et al. [108]. Compared to that work, a major change was the probabilistic modeling of the observed social signals. The authors took sensor noise into consideration and accounted for the possibility that connection events could be delayed or performed unintentionally. Another notable aspect of that model was the inclusion of context information, such as properties of the conversation topic or the interlocutors' present role (speaker or listener). The example use cases in that work were a collaborative placement task with a social robot and a training simulation for job interviews.

Processing

When somebody is busy processing information or deciding what to say, they usually avert their gaze to avoid distractions (see section 3.4.2). This behavior is interpreted very similarly when a robot displays it.

Andrist et al. [4] analyzed video recordings of human dyads and determined the probabilities for humans to look in a particular direction during cognitive activity. They found that humans looked mostly up (about 39.3%) or to the side (about 31.3%). The distribution is shown in figure 5.1. Gaze shifts according to this distribution were then implemented for the NAO robot and evaluated in a user study. The results showed that participants rated the robot's utterances as more thoughtful when its gaze indicated cognitive processing at the beginning of its turn than when it used static gaze or averted its gaze at inappropriate times.

Skantze et al. [122] found that humans were more likely to pause their drawing activity when the robot looked at the map during pauses in the instructions. While syntactically incomplete phrases had a slight inhibiting effect on their own, the difference only became noteworthy when combined with a gaze pattern that indicated the search for information on the map.

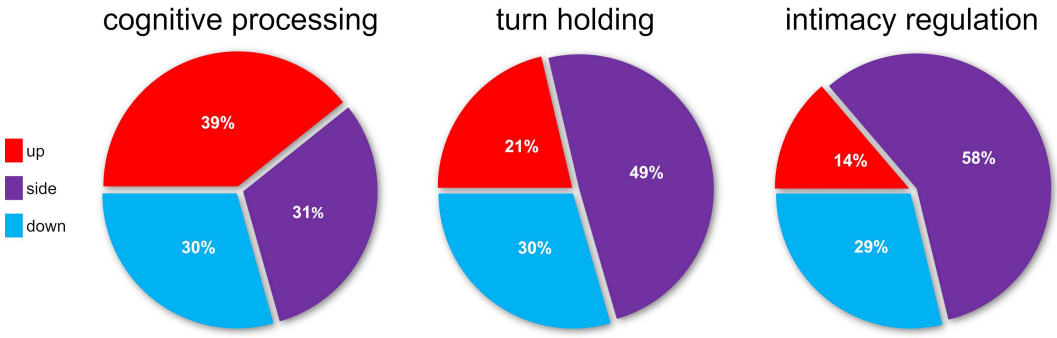


Figure 5.1: The distribution of gaze aversion directions based on the findings of Andrist et al. [4]. **Left:** Aversion during cognitive load. **Middle:** Aversion while holding the turn. **Right:** Aversion perceived as intimacy regulation.

5.3.2 Turn-Taking Intention

A dialogue system needs to understand the user's intentions regarding turn-taking and vice versa. As explained in section 3.4, overlapping speech does not necessarily mean that one interlocutor wants to interrupt the other, and neither does silence always mean that the speaker has finished their turn. The use of additional modalities, such as intonation or gaze, has proven helpful in human-agent communication.

Taking the Turn

One important task for a dialogue system is to distinguish between background noise, backchannel comments, and genuine attempts at interrupting the agent, also known as "barge-in".

Crook et al. [36] used both the intensity of the audio signal and the amount of time that the user spent speaking simultaneously with the agent. To count as an interruption attempt, the speech activity needed to either surpass a given intensity threshold or be of medium intensity for longer than a given duration threshold.

Selfridge et al. [119] relied on the semantic parsing of the overlapping audio for identifying valid barge-in attempts. They argued that speech acts should be ignored if they did not advance the dialogue, even if they were directed at the system. When a potential input was detected, their dialogue system used the speech recognizer's confidence measures to calculate whether its interpretation was likely to change. If this "stability" score exceeded a certain threshold, the system paused its output and waited for the assumed barge-in to finish. If the score dropped below the threshold again, the system resumed its output after a specified time-out. Once the speech recognizer provided a final result, the

dialogue manager tried to process and respond to the associated speech act. If it could not find a suitable response, the original output was resumed.

Chao [26] used both the voice activity and the recognized speech act to determine the user's intention. If the pitch of the audio signal exceeded a given threshold for a specified time, this was interpreted as the user trying to start speaking. When partial results were available from the speech recognizer, the user was assumed to have taken and be currently holding their turn. Since the speech channel was modeled as a limited resource, conflicts were only tolerated for a specified amount of time, and a duration threshold determined when the agent would yield the resource to the user.

Müller et al. [93] explored the use of SVMs for predicting the next speaker from the video signal of group conversations. By using head pose, eye gaze direction, parting the lips, and dropping the jaw, they already achieved the highest accuracy, and adding dynamic difference features did not improve the prediction. Furthermore, they observed that the accuracy was higher when features from all participants were considered, from which they concluded that an individual's behavior had to be interpreted in the context of the others' actions. However, their best results were still relatively inaccurate, with 0.51 to 0.60 recall for the trained classifiers compared to 0.50 for the baseline heuristic. Possible conclusions from this work are that the video signal alone is insufficient for predicting turn-taking behavior but temporal difference features are not necessary.

Holding the Turn

As mentioned in section 3.4.2, speakers pay attention to the listener when they expect feedback but avoid looking at them when they do not wish to be interrupted. This pattern was among those that Andrist et al. identified in their video analysis of human dyads [4]. According to their data, humans mostly look to the side (49%) or down (30%) when holding the turn. This distribution is also shown in figure 5.1. The user study that they conducted with a NAO robot confirmed that participants were more hesitant to interrupt it when it showed this gaze pattern during pauses. On average, they waited 608 milliseconds when the gaze was averted at the right time, compared to 331 milliseconds for static gaze and 327 milliseconds for poorly timed aversions.

Yielding the Turn

Besides interpreting user speech, a dialogue system also needs to infer the meaning of periods of silence. In particular, it needs to understand whether the user is pausing for thought, expecting a response from the agent, or addressing

another person who may or may not be involved in the conversation themselves. One way to approach this challenge is by considering multiple modalities.

For example, in a quiz game kiosk developed by Bohus and Horvitz [18, 17], the virtual agent assumed that the floor was handed to the person whom the previous speaker had been addressing. Said addressee was inferred from a combination of acoustic and visual sensor data, most notably the origin of the current voice activity and the head orientation of all human participants. The authors refer to the latter as the "visual focus of attention". Although they use a statistical approach rather than explicit rules, this further supports the link between turn management and attention, as explained in section 3.4.

The same pattern can be applied in the other direction. Whenever the virtual quiz master was ready to pass the floor to a human participant, it looked towards the next addressee and raised its eyebrows [18]. Similarly, Holroyd et al. [54] implemented a fixed policy of initiating mutual facial gaze at the end of the robot's turn.

5.3.3 Affective Traits and Relationships

Nonverbal behavior often carries information about the relationship between interacting parties. As section 3.4.3 explained, both gaze behavior and overlapping speech can send messages regarding status or affiliation. Additionally, the direction of the head is associated with certain personality traits.

Personality

Ball and Breese [14] proposed using a Bayesian Network for relating temporary emotions and long-term personality traits to observable surface behaviors. In their work, they equate personality to the tendency towards taking on a particular interpersonal attitude. Emotions are represented by valence and arousal. While their example of a Bayesian model does not link emotions and personality directly, the probability for observing a given behavior feature depends on a combination of those internal states. The authors argue that a Bayesian Network is well-suited for representing the inherent uncertainty in relating behavior to affective states and depicting these relationships in a way that is intuitively understood by humans. Moreover, they point out that these models are capable of both predictive and diagnostic reasoning, meaning that the same network can be used for generating a computer-controlled agent's behavior and inferring the affective state of a human user.

André et al. explored different approaches for modeling the personality and emotions of socially interactive agents [2]. They focused on extraversion, agreeableness, and neuroticism since those traits have the strongest impact

on interpersonal behaviors (see section 3.2.4). The examples they presented showed the following key aspects:

- **Separation between behavior selection and execution:** The affective reasoning components prepared action commands for the agents, but the actual playback of animations and sounds was left to the game engine that displayed the virtual environment.
- **Reactive behavior generation:** Agents responded directly to events that they perceived via (virtual) sensors. Those triggered the expression of *primary* emotions that are associated with physiological reactions, such as fear.
- **Deliberative behavior generation:** Agents actively pursued goals and evaluated the associated risks, successes, or the effects that various other events had on their goals. Such cognitive reasoning, based on the OCC model [98], was also required for triggering *secondary* emotions and, in particular, the selection of emotionally colored speech acts.
- **Playful exploration:** The authors observed that varying the underlying personality traits allowed for quick behavior adjustments. This, in turn, encouraged users to experiment with different combinations of character personalities.

Arellano et al. [6] showed that the direction of an agent's head influences what personality a human observer assigns to them. Their study used static images of a 3D character and had participants rate those according to the Big Five personality traits. According to their results, directing the head straight up or to the upper side makes the agent appear the most extroverted, whereas directing it straight down yielded the lowest extraversion rating. Directing it straight up was perceived as least agreeable, whereas directing it straight down was perceived as most agreeable. Finally, directing it straight up was rated as most neurotic, whereas directing it to the side was rated as most emotionally stable.

Besides body language, there are findings regarding speech timing and personality perception. Ter Maat et al. [130] implemented a conversation simulator that generated unintelligible dialogue audio, using both a randomly speaking agent and one that was scripted to employ specific strategies for starting and stopping its contributions. They then combined the respective strategies and had study participants rate the scripted agent's personality. According to their results, starting before the active speaker finished their sentence was perceived as less friendly, less agreeable, less respectful, and less

warm than speaking directly afterward or leaving a gap of a few seconds. Starting early was also perceived as more active and more dominant than starting after a gap.

Interpersonal Attitude

The gaze direction of agents was also examined in the context of interpersonal attitude. Fukayama et al. [43] displayed moving eyes to study participants, accompanied by a pre-recorded question-answering dialogue between the agent and a human. The animations varied with regard to several parameters and were interpreted as follows.

- **Amount of Gaze:** A gaze ratio of 50% was rated as mostly neutral in the "friendly"and "dominant" dimensions, whereas continuous staring was perceived as less friendly and more dominant. Deviating in the other direction, making eye contact only 25% of the time, made the agent seem less dominant.
- **Mean Duration of Gaze:** The default length of mutual gaze, set to 1000 milliseconds, was rated as mostly neutral regarding dominance and friendliness. Holding the gaze for half that time was perceived as less dominant. While no significance measure was reported for the other direction, the results indicate that holding it for 2000 milliseconds was perceived as less friendly.
- **Aversion Direction:** Looking down between phases of eye contact was perceived as least dominant compared to looking up or to the side. No significance measure was reported for the effect on friendliness, but the results hint at sideways aversion being perceived as less friendly.

Chollet, Ochs, and Pelachaud [28] extracted behavior sequences from a corpus of job interviews that had been annotated with the recruiter's interpersonal attitude. In a first step, they extracted those behaviors from the corpus that co-occurred with a perceived shift in attitude. They then trained a Bayesian network for generating behavior sequences that would express the desired combination of affiliation and status. After a separate module matched the agent's spoken message to appropriate nonverbal behavior, such as iconic gestures to complement descriptions, that Bayesian network was used to fill the gaps with additional gestures, gaze shifts, or facial expressions. Their evaluation showed that decreases in friendliness and increases in dominance were perceived as intended, whereas submissive signals were considered similar to the neutral baseline behavior.

As for speech activity, Ravenet et al. [106] used the agent's interpersonal attitude to calculate both the delay after which it would want to speak again and to decide whether it would talk over the current speaker(s). The delay was defined so that a submissive attitude would increase the delay, a dominant one would shorten it, and a friendly one would amplify that effect while a hostile one would not. As for the decision on speech overlaps, they considered the average attitude towards the current speaker(s). Their implementation was based on the idea that, while a submissive attitude would suppress speech and a dominant one would activate it, both hostility and friendliness increased the willingness to speak at the same time as the other person. The perception study they conducted with this system confirmed that the generated speaking behavior expressed the desired level of dominance and friendliness.

Glas et al. [47] combined varying amounts of overlap with different types of disruptive and cooperative speech acts. In their study, they presented the stimuli as videos of static humanoid silhouettes with subtitles, playing the speakers' voices on the left respectively right audio channel. Both agents appeared male and used exactly the same synthetic voice to control for biases. Their results showed the following patterns for the different interrupting acts.

- **Asking Questions:** The longer the interrupted agent continued speaking, the more dominant and less friendly it appeared. As for the interrupting agent, it appeared less dominant when it waited for a pause in the speaker's utterance. On-topic questions were perceived as more friendly, more engaged, and more involved than off-topic ones. The latter appeared less dominant, more friendly and more engaged when overlap was avoided.

- **Expressing Opinions:** Yielding the turn when the interrupting agent expressed a compatible opinion during a pause made the original speaker appear less friendly. Agreeing made the interrupting agent appear more friendly, more engaged, and more involved.

- **Managing the Partner's Communication:** The original speaker was perceived as more dominant when the interrupter completed the other's utterance rather than cutting it short. The interrupter was perceived as more dominant when it made the other stop quickly compared to when it waited for a pause. It appeared more friendly, more engaged, and more involved when it uttered a completion, and the difference in friendliness was greater when overlap was avoided.

- **Managing the Topic:** Interrupting during a pause made the second agent appear more friendly than overlapping with the other's speech,

regardless of whether it added information to the current topic or introduced a new one.

5.4 Adapting To The User

Personalization and adaptation are popular topics when it comes to user experience. In reality, it is hard to find a solution that suits everyone, and even when a system is developed with different personas in mind, there will always be some users who do not fit neatly into these categories.

Furthermore, psychological findings regarding compatibility and preferences in social relationships are rarely straightforward to implement. For example, a study by Mehlman [82] failed to confirm that friends are more similar than enemies. Tett and Murphy [131] found that people preferred co-workers whose personality allowed them to express their own personality traits, providing an alternative explanation for why certain people appear to collaborate more effectively. However, as Argyle and Little explained it [11], the apparent personality of a human can change drastically depending on context factors, such as the present observers or the role that this person is expected to play.

This section will present several works that examine the effects of matching users to different agent personalities. Afterward, it will look at approaches for doing so automatically.

5.4.1 Effects of Agent Personality

Tailoring a computer system to the user's preferences is expected to make it more acceptable and, consequently, more effective. There are several theories regarding compatibility, the most prominent ones being *similarity attraction* ("birds of a feather flock together") and *complementarity attraction* ("opposites attract"). Consequently, similarity and dissimilarity are frequently researched in human-computer interaction.

Three main topics have emerged while reviewing related works. One is *engagement*, the degree to which the user is interested and involved in the interaction. *Trust* in the agent determines whether the user would accept its help and feel safe using that system. A closely related concept is *persuasiveness*, the agent's ability to make the user follow its recommendations.

Engagement

Andrist et al. [3] varied the gaze behavior of a humanoid robot to express two different levels of Extraversion. Based on their observation of human dyads, they implemented different durations for looking at either the partner or a

Towers of Hanoi puzzle that said partner was supposed to solve. Furthermore, they found that the gaze durations changed depending on whether both parties were collaborating to solve the puzzle or discussing the task between puzzles. The normal distributions from which they sampled the gaze durations are displayed in table 5.1.

Interaction Type	**Gaze Target**	**Introverted**		**Extraverted**	
		M	SD	M	SD
Collaboration	partner	0.57s	0.19s	2.66s	0.80s
	workspace	11.65s	11.17s	4.04s	2.12s
Discussion	partner	1.59s	0.39s	3.91s	1.22s
	workspace	6.21s	8.14s	1.01s	1.26s

Table 5.1: Gaze duration distributions according to Andrist et al. [3], displaying either an introverted or an extroverted robot personality.

After confirming that the gaze patterns were perceived as intended, they systematically combined the two robot personalities with human study participants who scored either high or low on the Extraversion questionnaire. Participants were instructed to solve as many puzzles as they liked with the help of the robot, to test how the robot's personality influenced their motivation to collaborate with it.

The results showed that participants whose Extraversion level was the same as the robot's collaborated for longer in total, but only when they had no intrinsic motivation to solve the puzzles. No significant difference was found regarding their perception of the robot's performance.

Trust

Zhang et al. [143] examined whether similarities in the Big Five personality traits made drivers of automated vehicles perceive them as safer. The vehicle's personality was varied by combining normal respectively aggressive driving behavior with either sunny or snowy weather. Study participants were to rate their own personality before watching the four videos in random order and rating the personality and safety of the vehicle in each of them.

- **Agreeableness:** Safety scores were highest when both the participant and the vehicle scored high on Agreeableness. It was lowest when the participant was disagreeable and the vehicle appeared to be agreeable.
- **Conscientiousness:** Safety scores were highest when the participant scored low on Conscientiousness and the vehicle appeared to be highly

conscientious. The lowest rating was given when their Conscientiousness levels were the opposite.

- **Emotional Stability:** Safety scores were highest when the vehicle appeared to be emotionally stable, regardless of whether the participant scored low or high on this trait.

Braun et al. [19] ran a real-world study with different in-car voice assistant personalities. They had participants drive a predefined route and perform certain tasks along the way, once with a neutral assistant personality and once with a more specific personality. Assistant personalities were defined as combinations of "casual versus serious" and "equal versus subordinate". However, the work in question gives no concrete rules for matching the four personalities to the Big Five traits of the participants. Instead, (mis)match with the assistant is defined as whether the assigned personality is the same that the participants would have picked if given the choice.

Results showed that, compared to the neutral assistant, the matching assistant was perceived as more trustworthy and more likable. In contrast, the trend was reversed when the participants were paired with a different assistant personality than the one they preferred. In the latter condition, the mismatched assistant was rated less useful and less satisfying than the default one.

Persuasiveness

Moon [89] varied the message style of a computer system's recommendations to express different levels of dominance. Dominant messages used assertions and commands along with a fictional confidence level of 80% on average. In contrast, submissive messages combined questions and suggestions with an average confidence level of 30%. These message styles were then compared in two experiments.

The first experiment tested how likely participants were to change their ranking of cars. The results showed that recommendations with a dominant style more easily persuaded participants who scored high on dominance themselves. They also rated the information quality higher than submissive participants did in that condition. As for the system's expertise, participants whose dominance level matched the system's message style perceived it to be greater. In the second experiment, the computer presented news articles, classical music samples, cartoon recommendations, and health tips using the same variation of the message style. The results showed that participants rated all recommendations more favorably when they were presented in a style matching their own dominance level. Additionally, participants with a high dominance

score perceived the dominant system as more competent than the submissive participants did.

Stress

Gebhard et al. [45] compared two different agent personalities in the context of job interview training. They examined whether the virtual recruiter's behavior affected the subjective evaluation and the objectively measurable behavior of a human roleplaying as a job candidate. The two recruiter personalities, "understanding" respectively "demanding", were designed as follows:

- **Speech timing:** The *demanding* agent uses longer pauses within its turns, which the authors explain as "show[ing] dominance in explanations and questions". While the authors do not elaborate on why this behavior would appear dominant, it could be interpreted as the agent holding the floor without actually needing it.
- **Gaze direction:** The *understanding* agent spends more time looking at the user while speaking (3000 to 5000 ms for eye contact versus 500 to 1000 ms of aversion). Furthermore, it does not avert its gaze while listening, whereas the *demanding* agent occasionally does.
- **Linguistic style:** The *understanding* agent frequently uses polite, appreciative wording, whereas the *demanding* agent rarely does.
- **Facial expressions:** The *understanding* agent display positive emotions, whereas its *demanding* counterpart expresses negative ones.
- **Gestures:** The *demanding* recruiter occupies more space with its gestures compared to the *understanding* one. Additionally, the latter tends to tilt the head sideways.

Study participants interacted with both agents in random order while their movements, facial expressions, and a wide range of audio features were recorded. The analysis of those recordings showed that users produced shorter utterances and made more breathing pauses when interacting with the *understanding* agent, from which the authors concluded that the participants were feeling less pressure to answer quickly. Self-reported measures confirmed that the participants felt more comfortable, less stressed, and less challenged in that condition. They also perceived the *demanding* agent as less natural, which Gebhard et al. explained with the expectation that a recruiter should be "friendly and supportive".

5.4.2 Adaptation Approaches

In reality, it is rarely possible to find one solution that works for all users. Although a neutral agent personality may be better than a mismatched one, as mentioned above, there is often room for improvement via tailoring the agent to a specific user. In other words, the closer an agent is to the user's preferred archetype, the better they appear to collaborate with them.

There are two main approaches for adapting an agent's personality to the target user. One way would be to configure it explicitly, for example, based on theoretical compatibility with user traits or according to the user's preferences. The other would be to have the agent adapt autonomously, based on some success metric for the interaction. The latter is especially useful when users might not know their preferences or when there is a conflict between what they like and what actually motivates or convinces them.

Selection of Archetypes

Several studies that examine compatibility between human and agent personality start by first asking the human to rate their own personality. For example, Moon [89] had participants rate themselves regarding sex roles, arguing that the questionnaire used for this had been found to correlate with interpersonal dominance measures. People were sorted into either the "submissive" or "dominant" group, depending on whether they scored below or above the median value. For the experiments, they were matched with a computer system that used either a dominant or a submissive message style to compare pairings that were either similar or opposite in this dimension.

Andrist et al. [3] used a similar method, having people rate themselves on the Extraversion scales of the Big Five Inventory. Each item was measured on a five-point Likert scale, and people scoring lower than 2.5 on average were classified as "introverted", whereas those scoring higher formed the "extroverted" group. These classifications were first used to observe the gaze behavior of humans belonging to either level and transfer these behavior patterns to a humanoid robot. In the experiment, participants were paired with a robot exhibiting either the same or the opposite level of Extraversion.

Zhang et al. [143] used a shorter questionnaire, the Ten Item Personality Inventory [49], to measure all Big Five traits. Participants filled it in for their own personality as well as the one they perceived from the automated vehicle in each experimental stimulus. The scores for each trait were classified as "low" or "high" compared to the mean. When analyzing the ratings for the vehicle's perceived safety, the authors considered whether the participant had scored high or low on a given trait and whether they perceived the vehicle as similar or dissimilar regarding that trait.

It should be noted that while compatibility is often modeled in terms of similarity, it is also possible that more complex rules determine a user's preference for an agent. Braun et al. [19] prepared four agent personalities based on the possible combinations of "casual" versus "formal" and "equal" versus "subordinate". They then let 31 participants rate their own personality in terms of the Big Five traits and choose their preferred personality for an in-car voice assistant. The answers were used to construct a decision tree assigning a matching agent to participants in the following compatibility study. However, it turned out that only 16 out of the 55 study participants would have chosen the personality that the decision tree assigned to them. The authors concluded that the training sample was too small and suggested that it would be more realistic to let users explicitly select their preferred assistant.

Online Adaptation

Manually selecting an agent personality may not always be possible or desirable. As Braun et al. pointed out, users might ignore personalization options if the default is good enough, missing out on the possible improvement [19].

As for automatic matching, there is the problem that theories regarding compatibility contradict each other. Sometimes people might prefer similar personalities and attitudes [89, 55, 3, 143], and sometimes they might prefer complementary ones [55, 120, 143].

A possible solution to these issues could be to have the agent adapt automatically. For example, Ritschel et al. [110] proposed using reinforcement learning based on the user's social signals. A Bayesian network was used to relate observable behaviors, such as leaning forward or looking away, to the user's engagement during the interaction. The robot could choose between increasing its Extraversion level, decreasing it, or leaving it unchanged. The selected action was then rewarded or punished depending on how the user's engagement changed in response.

While personality is rarely adapted directly, the term "adaptation" is commonly used in the context of empathy. For example, Conati referred to a context-sensitive virtual butler as an "adaptive user interface" [35] due to its ability to take the context into account and avoid negatively impacting the user's mood. Leite et al. [77] considered adaptation to the user's affective state an important social skill for a robotic chess tutor interacting with children. In their application, the iCat robot had different empathic strategies at its disposal and employed reinforcement learning to determine which strategies were most effective in improving a given user's mood.

Neither of these works mentions an explicit agent personality. However, they indicate an underlying pattern of sharing the user's goals that, in turn,

hints at a high degree of affiliation between both interacting parties (see section 3.2.3). In other words, the better an agent manages to account for the human's goals, the more it appears affiliated with the latter.

5.5 Conclusion

Many researchers have explored ways to optimize the turn-taking of ECAs, either by deducing the user's intention or by clearly displaying that of the agent. However, the agent's personality is rarely modeled explicitly. Most works focusing on action timing try to achieve idealized, compliant behavior to optimize efficiency, user affect, or overall engagement. In many cases, overlaps and interruptions are seen as mistakes that need to be avoided, but occasionally, the potential of allowing them has been explored.

Few researchers have used Bayesian networks to represent the probabilistic relationship between the agent's nonverbal behaviors and personality [14] or interpersonal attitude [28]. However, none of those works modeled the agent's underlying goals to explain why one behavior was more appropriate than others. So far, such models have only been trained on statistical co-occurrence, so the utility-based behavior selection is a major aspect that this thesis will add.

5.5.1 Action Timing

One useful core concept is the "minimum necessary information" (MNI for short) [26]. Any approach for semantically plausible interruptions must take into account whether the agent has heard enough to proceed. In simple scenarios, such as a short conversation with a limited domain vocabulary, the end of the MNI can be marked by hand. However, for more complex scenarios, it would be necessary to look for syntactic completeness or the confidence that the NLU component has in parsing an appropriate input. The MNI could, for example, consist of a subject and predicate (who does what) or any single word that unambiguously identifies the user's intention (such as a unique color in a phrase like "hand me the *blue* book, please").

Once the agent has heard enough to act on it, the question remains for how long it should continue to listen. It is possible to model this as a resource management problem, requiring the agent to wait for several prerequisites, such as an unoccupied voice channel and the visual attention of the interlocutor.

Another component of most approaches is a set of timeouts, for example, to detect when accessing a resource has failed. Most of those can be varied to achieve different personality impressions. However, the appropriate timing

may vary between contexts, depending on factors like the user's mood and current intentions.

The latter cannot be observed directly, so the system needs to take various forms of uncertainty into account. Decision-theoretic approaches can be employed to handle such uncertainties on various levels, from sensor noise and output latencies to non-deterministic or ambiguous user behavior.

Consequently, this thesis will focus on a **decision-theoretic approach**. The completion of the MNI, resources such as the **visual attention**, or the passing of certain **time thresholds** will be core observations from which the system will draw conclusions about the appropriate agent response.

5.5.2 Behavior Reflecting Internal States

Many communicative behaviors have been successfully transferred from human dyads to human-agent interaction. Among the most basic and, at the same time, most important patterns is signaling attention by turning the head or eyes towards what is relevant in this moment. Visual attention is closely linked to turn-taking intentions and has been studied in this context, confirming that humans take it into account when interacting with social robots. Furthermore, the direction of the human's gaze plays a role in measuring engagement as an indicator of the interaction quality.

In general, multimodal signals help disambiguate turn-taking intentions. Besides gaze, the intensity and duration of speech overlaps have already been used to distinguish between backchannel comments and genuine barge-in attempts (see section 5.3.2). However, nonverbal behaviors also influence how humans perceive an agent's interpersonal attitude and, by extension, its underlying personality. The actual semantics of the message also greatly influence how speech overlaps are perceived. Consequently, many factors have to be considered at the same time.

This thesis will focus on the **gaze behavior** of both participants in combination with **overlap and silence durations**.

5.5.3 Adaptation

Several studies have examined how the personalities of humans and agents affect their interaction. There is evidence for similarity attraction on some traits and for complementarity attraction on others. Overall, it appears that an agent that meets the user's preferred personality will be more effective in achieving the application's goals.

However, only some works have explored automatically assigning a compatible agent personality to a given user. The rules for which personality is

preferred under which circumstances are not straightforward, so a possible solution would be to adjust the agent's personality traits gradually. However, this has rarely been done. Instead, research has focused on selecting agent behavior (such as empathic reactions or action timing) that directly optimizes a measure of the interaction quality, such as the user's affective state.

This thesis intends to **expose parameters** to interaction designers from which behavior patterns can be derived systematically. While online adaptive systems are outside this work's scope, it will contain a brief discussion of how these parameters could be adjusted automatically.

Part II

Approach

Chapter 6

The Turn-Taking Model

6.1 Introduction

While interacting with other people, humans keep both their own goals and those of the other party in mind. How much the latter influence a person's behavior may depend on a wide range of factors - how empathetic they are, how much they want to please the other, or how much they fear the consequences of standing up for their own goals. At the same time, reasoning about those goals requires reasoning about aspects that are not easy to predict. Humans cannot read each other's minds, so they can never be certain what the other party is planning, what they may actually do, or how they may react to events.

Human reasoning often involves "following their gut feeling" or acting on flawed but easily applicable heuristics [123, 1]. Such approaches are not easy to implement on logical machines. Since there are no reliable rules for the machine to follow, it is hard to program them traditionally and prepare them for all possible situations. A strictly logic-based model, such as the structured actions and belief updates described by Cohen and Levesque [34], would soon become too complex for maintenance and extension, especially when the system needs to accommodate a diverse user base.

Statistical approaches - those that are currently labeled as *Artifical Intelligence* - are being heralded as the solution to such problems, with their ability to discover unwritten rules from recurring patterns in the dataset. However, those approaches come with their own downsides. Demands for the transparency and accountability of these technologies become louder and louder, and there is an entire research field dedicated to making Artificial Intelligence explain its decisions to domain experts or naive end customers.

To further complicate matters, humans have been found to apply differ-

ent criteria when judging a machine's action than when judging a human [78]. While a "cold" rational choice is often frowned upon in humans, that rationality tends to be expected from artificial agents. In fact, many humans are alienated by the very idea of a machine acting on emotions, despite the widespread phenomenon of humans subconsciously interpreting an ECA's behavior as human-like thought patterns and intentions. (See chapter 5 for examples of behaviors that are interpreted that way.)

A plausible compromise is to give an agent human-like reasoning but to implement it in an idealized manner. Therefore, this thesis builds upon decision theory [123, 1] as a structured way of balancing the costs and benefits of the agent's actions against each other. As shown in chapter 5, methods such as *Bayesian networks* [14, 56] or *dynamic decision networks* [35] have already been applied or proposed for similar use in human-agent interaction.

The model for the agent's behavior was developed around the following requirements.

- **Decision-theoretic Behavior Selection:** The agent needs to choose the conversational timing and gaze target that are most likely to help it achieve its interaction goals. Uncertainty regarding the other participant's intentions should be taken into account.

- **Personality-based Interaction Goals:** The agent's goals need to align with the personality that a human observer should attribute to it. The relationship between personality traits and interaction goals should be grounded in psychological literature and artistic conventions so that the model remains transparent for behavior designers.

- **Domain-agnostic Interaction Goals:** The model needs to be reusable for various interaction scenarios. Goals should be as abstract as possible while allowing concrete action choices.

This chapter describes the concepts behind the turn-taking model in the form that evolved alongside the practical applications described in part III. First, it will explain the structure of the influence diagram and how the different aspects, such as personality, cognitive states, and communicative goals, are represented. After that, it will look at the probability distributions used in the final prototype and explain how they were chosen or calculated. The chapter ends with a summary.

6.2 Network Structure

The purpose of this influence diagram is to represent the connection between the agent's affective model, its cognitive state, its actions, and their impact on its goals. Figure 6.1 shows an overview of the final behavior model. Its evolution will be discussed in more detail in chapter 10.

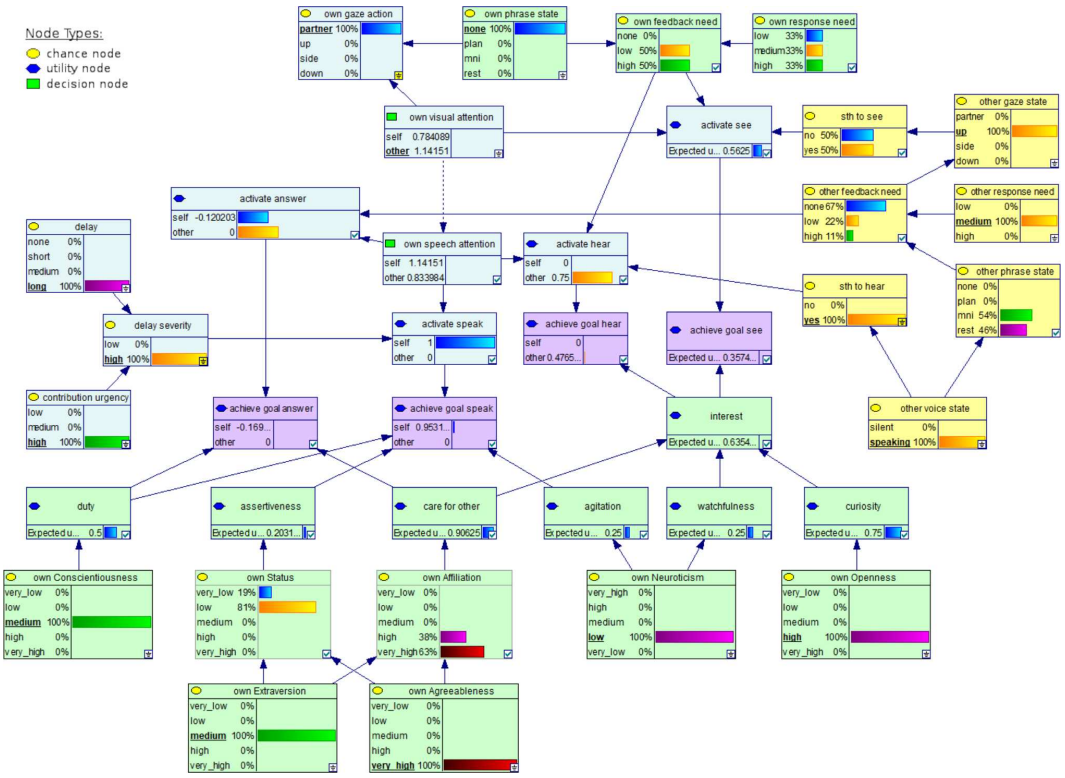


Figure 6.1: The final revision of the turn-taking model for the interactive prototype.

6.2.1 Affect

The agent's affect model consists of its personality and interpersonal attitude. Other aspects, such as the current emotional state, were explored during this thesis. However, they did not bring an improvement worthy of the added complexity, so they were ultimately dropped from the model. The same goes for the three-dimensional attitude models examined in section 3.2.4. While the findings reviewed in that section appear useful in general, it was hard to apply them in practice.

Personality

Chance nodes can not only describe probabilistic outcomes but can also be used to represent input parameters by "observing" their desired value. This allows the interaction designer to configure the agent's personality and, consequently, its primary behavior.

The Five Factor Model was chosen because it is compact but expressive, and a large body of research connects it to behavior or describes its use in computer science. Consequently, the influence diagram contains five nodes representing the different personality traits. For this thesis, they are discretized into five levels each, ranging from a very low expression of this trait to a very high one.

Interpersonal Attitude

As explained in section 3.2.3, interpersonal dynamics can be described using either personality traits or the interpersonal attitude dimensions. In particular, *Extraversion* and *Agreeableness* are an alternative set of axes for defining the interpersonal circumplex [80, 39]. Personality is argued to change slowly and remain mostly constant throughout a person's life [44], which implies that their default attitude towards people can be derived from these traits.

However, Argyle and Little [11] point out that this seemingly constant personality depends greatly on the situations in which a person is being observed, as well as the relationship to the present observers and the role the person is supposed to act out. For instance, a person may act reserved in front of their boss and colleagues at work but be more relaxed and outgoing when visiting a bar with friends. They may be cold and distrustful towards strangers but open and loving when interacting with close family members.

Therefore, it seems reasonable to separate the intrinsic personality traits from the context-dependent interpersonal attitude. Modeling them in separate layers opens up the possibility of modifying the default attitude in future work by connecting it to other influence factors than the agent's personality traits. For example, one could imagine deriving the affiliation from the amount of time that the agent has known the present users or the status from the role that it takes on in the current conversation.

6.2.2 Cognitive State

Humans constantly monitor whether their interaction partner is able and willing to listen [58] or give them more time to start speaking when it seems like they are searching for the right words [4, 122]. Therefore, a systematic

turn-taking model is impossible without taking these cognitive processes into account.

While the semantic information determines the higher-level dialogue flow, the cognitive state of the interacting parties regulates the fine-grained timing of individual utterances or gaze shifts. In this thesis, the cognitive states are part of an influence diagram that opens or closes the communicative channels at the right time.

Attention

Turn-taking is closely coupled with attention, and the gaze signals associated with coordinating the speaker roles are hard to separate from the search for information (see section 3.4.2). Therefore, attention emerged as a key aspect of the computational model in this thesis.

The agent's attention is split into separate channels to generate flexible multimodal behaviors. For example, the agent's acoustic attention can focus on its own contribution while its visual attention is on the feedback signals of the partner. The attention target determines the agent's action with regard to the respective modality.

The influence diagram is used to reason about the observed interaction context and the attention target that will best satisfy the agent's goals. After synchronizing the sensor inputs and the tracked situation parameters with the influence diagram, the expected utility of attending to the possible targets is inferred. The agent then switches its attention to the target that maximizes the utility for its interaction goals, for example, by turning its gaze to the side while continuing to speak.

Speech Desire

Authors like Goldberg [48] have suggested that certain contributions, such as a clarification request or a comment on the speaker's message, are more sensitive to delays than others. Some models for agent behavior, such as the one implemented by Ravenet et al. [106], also include the *desire to speak* as a factor.

Consequently, this thesis assumes that the desire to speak depends not only on the interaction goals but also on the communicative function of the exchanged messages. For example, an answer to the partner's question is considered more urgent than a reaction to their feedback. In the presented turn-taking model, this is expressed as the intrinsic *urgency* of the agent's own contribution and the *delay* that amplifies it.

Unfortunately, no literature was found that would provide concrete values for this intrinsic urgency. For this thesis, they were defined according to the

subjective impact that delaying or dropping such a speech act would have on a functional level, for example, by hindering the conversation from progressing.

Information Need

A closely related concept to cognitive load and contribution urgency is the interlocutors' need for information. When cognitive load is high, humans tend to avoid distractions unless they require additional information for the planning process. In the latter case, they usually focus on the object they are talking about rather than the partner's reactions [10]. However, when they are interested in the partner's processing of what was said, they tend to look at their face to see signs of emotions or confusion [8, p. 170,172].

In this work, the need for information is assumed to depend on the progress of the current contribution. For example, if the speaker has not finished the MNI, the listener is likely to seek more information while the speaker is likely to avoid it. Afterward, however, the speaker is likely to seek feedback about what they said.

Furthermore, certain communicative acts require a dedicated response, which overlaps with the contribution urgency mentioned before. Here, this *need for a response* is assumed to depend on the communicative function in a similar way as said urgency. For example, a request has a stronger need for a reply than an expression of acceptance does.

As with the basic urgency of certain speech acts, there was no literature mapping it to concrete values. It could possibly be approximated from the analysis of dialogue recordings by looking at the average frequency of a particular act receiving a response. For this thesis, however, the response need was defined based on whether a speech act with that communicative function typically serves as a response itself or asks for something to be done on the addressee's part.

6.2.3 Interaction Goals

The goals themselves are not defined directly but rather represented by the utility that certain actions have for them. Additional weight and activation parameters control how much they contribute to the overall expected utility of the agent's possible behaviors.

Goal Representation

Each interaction goal is realized as several nodes in the influence diagram. First, there are regular utility nodes that model the benefit that an action has for the goal in question. Second, there are re-purposed utility nodes that

express the weight of a certain goal depending on the agent's personality, according to the idea that different personalities assign different priorities to the same fundamental goals (see section 3.3.2). Similar utility nodes express the *activation* of the goal, based on the theory that not all goals are on a person's mind at all times (see section 3.3.2).

Finally, a multi-attribute utility node represents the actual utility for a given goal. These systematically combine the benefits, the weights, and the activation of that goal to calculate the utility that the agent's actions have for said goal. Based on these utilities, the agent can then select the actions that contribute most to its activated goals.

Actions Affecting Them

Attention to the different communicative channels is modeled separately, based on the idea that they require different resources [26]. For example, people can listen while having their eyes on the object that the speaker is describing or while watching the road that they are driving on. In contrast, listening to another person is difficult while the listener is speaking themselves.

There are two possible actions for each modality[1]: opening the channel for attending to the *other*, or closing it in order to focus on the *self*. On the verbal channel, this means that the agent will speak when the attention target is itself and be silent when the focus is on the other participant. As for gaze, opening the channel means looking at the partner, whereas closing it means averting the gaze. Table 6.1 gives examples of meanings that are typically associated with combinations of these actions.

		visual attention	
		self	**other**
verbal attention	**self**	planning or holding the turn	seeking feedback from the listener
	other	processing what the speaker said	paying full attention to the speaker

Table 6.1: Combinations of attention targets mapped to situations in which they are typically observed.

Secondary actions may depend on these basic decisions. For example, in this thesis, the probability distribution for looking at a certain target varies

[1]Note that further modalities could be added by inserting new decision nodes with only two actions each. If the attention were modeled in a single node, this would require $2^{\#modalities}$ distinct actions.

according to whether the agent tries to make eye contact or avoids the other's gaze.

Prospects

Whenever possible, the consequences of an action should be represented by concrete *prospects* (see section 4.2.2) that are grounded in findings and theories about human communication.

For example, the severity of delaying an utterance depends on the inherent urgency of the message and the duration of the delay. The amount of information gathered from listening to or watching the other person depends on what the latter is doing.

Depending on the available findings in the literature, it may still be necessary to define some prospects based on heuristics. The challenge is to find a good balance between a simple but understandable model and a complex but realistic one.

6.3 Network Parameters

The conditional probabilities in a Bayesian Network can be assigned manually or extracted from a suitable data corpus. The first approach can draw on existing behavior rules and dependencies that psychologists have already identified. The second one requires the recruitment of a diverse population, a suitably general interaction topic, the recording of sufficient amounts of material, and finally the labeling of the relevant observations. For this thesis, the theory-based approach was chosen, but the network parameters could certainly be improved by analyzing a suitable data corpus.

6.3.1 Gaze

Concrete gaze distributions are hard to find, especially since they strongly depend on the cultural background. Also, much literature on body language still refers to the research aggregated by Argyle and Cook in the 1970s [8]. It is easier to find percentages or time thresholds in computer science literature, but one must be aware of the population that proposed and evaluated them.

One important source for this model was the study by Andrist et al. [4] in which they annotated videos of human dyads to associate gaze aversion directions with different communicative purposes. The probabilities that they determined for gaze aversions were linked to the utterance progress as shown in table 6.2. To assign them, the following mental states were assumed during an utterance:

phrase state	none	planning	mni	rest
up	0.137	0.393	0.213	0.213
side	0.575	0.313	0.492	0.492
down	0.288	0.294	0.295	0.295

Table 6.2: Conditional probabilities of averting the gaze in a certain direction, based on Andrist et al. [4].

- **None:** While the agent is not speaking, it shows *intimacy regulation* behavior and breaks eye contact to avoid staring at the partner. This is the default state between the agent's turns.
- **Planning:** While the agent has taken the turn but is not yet ready to speak, it shows *cognitive processing* behavior and avoids distractions to focus on its contribution. In technical terms, this phase begins when the speech command is issued to the agent's behavior realizer.
- **MNI:** While the agent is speaking the meaningful part MNI of its sentence, it shows *turn holding* behavior to signal that it does not want to be interrupted. In technical terms, this phase begins when the agent's TTS output has started to play.
- **Rest:** While the agent is speaking the rest of its sentence, it continues to show *turn holding* behavior to signal that it does not want to be interrupted. In technical terms, this phase begins when the agent's TTS output has reached the bookmark that signifies the end of the MNI.

6.3.2 Interpersonal Attitude

In section 3.2.3, it was explained that the axes of the Interpersonal Circumplex and the two corresponding Big Five traits, Extraversion and Agreeableness, are rotated about 30-45° relative to each other [80, 39]. In other words, there is a deterministic interdependence between the values of the four variables.

To calculate the conditional probabilities for this interdependence, the levels of each personality trait variable were mapped to the numeric range of $[-1.0, 1.0]$. The range was split into five equally sized subranges, each providing four samples for possible personality configurations.

In the next step, Excel spreadsheets were used to calculate the Status and Affiliation values resulting from all 400 combinations of these samples. The coordinates on the *(Extraversion, Agreeableness)* plane were rotated by an angle of $\alpha = -37.5°$, which corresponds to the middle of the aforementioned angle range. The new coordinates were calculated as follows:

$$Affiliation = cos(\alpha) * Agreeableness - sin(\alpha) * Extraversion$$
$$Status = sin(\alpha) * Agreeableness + cos(\alpha) * Extraversion$$

Finally, the *(Affiliation, Status)* coordinates were discretized into the five levels according to the subranges specified earlier. These relative frequencies of observing a specific Affiliation or Status level, given any combination of the personality trait levels, were then used as the conditional probability distributions at the interpersonal attitude nodes.

The final distributions can be seen in tables 6.3 and 6.4. Detailed tables of the intermediate results can be found in appendix A.1.

The study that was later conducted with those parameters (described in section 9.4 later) confirmed that the agents' personalities and attitudes were perceived as expected. An angle of $-37.964\,49°$ was found to minimize the error between the predicted Status level and the ratings given by the study participants. Details on the calculations are presented in section 9.4.3, and that angle was used to refine the conditional probabilities for the interactive prototype.

6.3.3 Feedback Need

The need for feedback is derived from the current speech phase as well as the current need for a response. Unfortunately, it cannot be measured objectively, and so no concrete data was available for mapping the interlocutors' feedback need to the current dialogue state. Therefore, the conditional probabilities (see table 6.5) were chosen according to the following principles.

- **Readiness to listen:** Before taking the turn and after speaking the MNI, an interlocutor has the potential to pay attention to the other. Therefore, the feedback need equals the need for a response.

- **Avoiding distractions:** The utterance planning phase blocks the need for a response, resulting in no need for feedback.

- **Monitoring understanding:** While delivering the MNI, the speaker observes the listener's reactions but does not necessarily pay full attention. Therefore, the feedback need can be anywhere between none at all and the level of the response need.

Agreeableness:	v. disagreeable	disagreeable	neutral	agreeable	v. agreeable
Extraversion:	very introverted				
very hostile	1.0000	0.9333	0.1875	0.0000	0.0000
hostile	0.0000	0.0667	0.8125	0.3750	0.0000
neutral	0.0000	0.0000	0.0000	0.6250	0.6875
friendly	0.0000	0.0000	0.0000	0.0000	0.3125
very friendly	0.0000	0.0000	0.0000	0.0000	0.0000
Extraversion:	introverted				
very hostile	1.0000	0.3750	0.0000	0.0000	0.0000
hostile	0.0000	0.6250	0.6250	0.0000	0.0000
neutral	0.0000	0.0000	0.3750	0.8125	0.0625
friendly	0.0000	0.0000	0.0000	0.1875	0.8750
very friendly	0.0000	0.0000	0.0000	0.0000	0.0625
Extraversion:	neutral				
very hostile	0.6250	0.0000	0.0000	0.0000	0.0000
hostile	0.3750	0.8125	0.0625	0.0000	0.0000
neutral	0.0000	0.1875	0.8750	0.1875	0.0000
friendly	0.0000	0.0000	0.0625	0.8125	0.3750
very friendly	0.0000	0.0000	0.0000	0.0000	0.6250
Extraversion:	extraverted				
very hostile	0.0625	0.0000	0.0000	0.0000	0.0000
hostile	0.8750	0.1875	0.0000	0.0000	0.0000
neutral	0.0625	0.8125	0.3750	0.0000	0.0000
friendly	0.0000	0.0000	0.6250	0.6250	0.0000
very friendly	0.0000	0.0000	0.0000	0.3750	1.0000
Extraversion:	very extraverted				
very hostile	0.0000	0.0000	0.0000	0.0000	0.0000
hostile	0.3125	0.0000	0.0000	0.0000	0.0000
neutral	0.6875	0.6250	0.0000	0.0000	0.0000
friendly	0.0000	0.3750	0.8125	0.0667	0.0000
very friendly	0.0000	0.0000	0.1875	0.9333	1.0000

Table 6.3: Conditional probabilities of observing a given Affiliation level for the given personality trait configuration.

Agreeableness:	v. disagreeable	disagreeable	neutral	agreeable	v. agreeable
Extraversion:	very introverted				
v. submissive	0.0000	0.0625	0.6250	1.0000	1.0000
submissive	0.3125	0.8750	0.3750	0.0000	0.0000
neutral	0.6875	0.0625	0.0000	0.0000	0.0000
dominant	0.0000	0.0000	0.0000	0.0000	0.0000
very dominant	0.0000	0.0000	0.0000	0.0000	0.0000
Extraversion:	introverted				
v. submissive	0.0000	0.0000	0.0000	0.3750	0.9333
submissive	0.0000	0.1875	0.8125	0.6250	0.0667
neutral	0.6250	0.8125	0.1875	0.0000	0.0000
dominant	0.3750	0.0000	0.0000	0.0000	0.0000
very dominant	0.0000	0.0000	0.0000	0.0000	0.0000
Extraversion:	neutral				
v. submissive	0.0000	0.0000	0.0000	0.0000	0.1875
submissive	0.0000	0.0000	0.0625	0.6250	0.8125
neutral	0.0000	0.3750	0.8750	0.3750	0.0000
dominant	0.8125	0.6250	0.0625	0.0000	0.0000
very dominant	0.1875	0.0000	0.0000	0.0000	0.0000
Extraversion:	extraverted				
v. submissive	0.0000	0.0000	0.0000	0.0000	0.0000
submissive	0.0000	0.0000	0.0000	0.0000	0.3750
neutral	0.0000	0.0000	0.1875	0.8125	0.6250
dominant	0.0667	0.6250	0.8125	0.1875	0.0000
very dominant	0.9333	0.3750	0.0000	0.0000	0.0000
Extraversion:	very extraverted				
v. submissive	0.0000	0.0000	0.0000	0.0000	0.0000
submissive	0.0000	0.0000	0.0000	0.0000	0.0000
neutral	0.0000	0.0000	0.0000	0.0625	0.6875
dominant	0.0000	0.0000	0.3750	0.8750	0.3125
very dominant	1.0000	1.0000	0.6250	0.0625	0.0000

Table 6.4: Conditional probabilities of observing a given Status level for the given personality trait configuration.

phrase state	none		
response need	low	medium	high
none	0.00	0.00	0.00
low	1.00	0.50	0.00
high	0.00	0.50	1.00
phrase state	plan		
response need	low	medium	high
none	1.00	1.00	1.00
low	0.00	0.00	0.00
high	0.00	0.00	0.00
phrase state	mni		
response need	low	medium	high
none	1.00	0.50	0.33
low	0.00	0.50	0.33
high	0.00	0.00	0.33
phrase state	rest		
response need	low	medium	high
none	0.00	0.00	0.00
low	1.00	0.50	0.00
high	0.00	0.50	1.00

Table 6.5: Conditional probabilities of requiring feedback, based on the current utterance progress and the fundamental response need for the uttered speech act.

6.3.4 Contribution Delay Severity

Like the need for feedback, the inherent urgency of the pending contribution and the delay duration thresholds are defined outside the behavior model. They combine to form the *delay severity* according to the heuristic in table 6.6.

6.4 Conclusion

This chapter presented a concept for an influence diagram that chooses the agent's target of attention to maximize the benefit for its communicative goals.

delay	none		
contribution urgency	low	medium	high
low	1.00	1.00	0.75
high	0.00	0.00	0.25
delay	short		
contribution urgency	low	medium	high
low	1.00	0.75	0.50
high	0.00	0.25	0.50
delay	medium		
contribution urgency	low	medium	high
low	0.75	0.50	0.25
high	0.25	0.50	0.75
delay	long		
contribution urgency	low	medium	high
low	0.50	0.25	0.00
high	0.50	0.75	1.00

Table 6.6: Conditional probabilities of the delay severity, based on the urgency level of the utterance and its current delay.

A cluster of chance nodes represents the agent's personality, allowing the designer to configure its traits by "observing" the desired level. The agent's interpersonal attitude is derived from Extraversion and Agreeableness, based on their connection in psychological literature and the idea that personality defines someone's default attitude toward unknown people.

The agent's personality traits and interpersonal attitude determine how the available goals are weighted, so they contribute differently to the expected utility. The conversation state determines whether the goals are active in the first place. Finally, the attention target determines the direction in which information can flow and, consequently, the degree to which the respective goals can be fulfilled.

Attention targets are modeled separately for each modality. This separation allows for human-like flexibility in choosing the source of information or switching to a free channel for transmitting a message. As for the exchanged information, the need to send or receive it depends on the communicative function of what is said. It is not built into the influence diagram but rather set as an observation whenever the dialogue progresses.

The next chapter will explain how the model can be embedded in a dialogue application. It will provide more details on the knowledge that needs to be observed by the influence diagram and show how its decisions are used to regulate the agent's turn-taking behavior.

Chapter 7

The Participant Framework

7.1 Introduction

A decision-theoretic model cannot work on its own. To evaluate the proposed turn-taking model, it also had to be embedded in at least one working dialogue application.

The first step was to test it with two computer-controlled agents as a proof of concept. They were implemented as two separate processes without direct access to each other's mind states so that they could simulate the barrier between a human user and an ECA. The idea between this artificial separation was that one of these agents would eventually be replaced by a human whose intentions could only be inferred from their surface behavior. Details on these two setups will follow in chapters 9 and 10.

Consequently, a structured way to connect different participant types was required. The agents needed the ability to perceive both their own kind and the human in front of their sensors, as well as make sense of the raw input. They needed to pass their observations on to the influence diagram and monitor it for changes in its decisions. Finally, they needed to use these decisions to exhibit the associated behavior.

The behavior model was intended to work with arbitrary participant constellations, such as human-agent dyads, a rule-based agent with a learning agent, or any number of agents with any number of users. This lead to the following requirements for the surrounding dialogue application.

- **Agent-agnostic Behavior Definition:** The modeled behaviors need not only be separated from the intended meaning but also from their technical realization via the agent's hardware and software.

- **Messaging Infrastructure:** Knowledge needs to be distributed systematically. This, in turn, requires standardized semantic representations and a suitable infrastructure for exchanging information between participants.

- **Extensible Agents:** Common design patterns need to be identified to model a wide range of participants, from human users to rule-based ECAs and those augmented with machine learning.

This chapter will explain how the agents represent and exchange semantic knowledge. It will look at how messages are standardized, tied to the end of the MNI, transmitted, and retrieved from an agent's memory. Afterward it will describe the connection between the participants and the rest of the setup.

7.2 Knowledge Representation

Certain behavior patterns only make sense when one considers the context. Sections 3.4 and 5.3.3 showed that humans have different tolerances for cooperative speech overlaps than for domineering ones and that a response only counts as such if it is semantically linked to what came before. An action may need to be postponed if it would impact the user's mood negatively, and a clarification question may need to be asked early to minimize the risk of misunderstandings. Finally, a meaningful response is not possible before one knows what to respond to.

Therefore, turn-taking needs to consider not only who is speaking but also what has been said so far. The same goes for other modalities, such as gaze signals or raw voice activity.

Figure 7.1 illustrates the two types of knowledge managed by an agent participant.

The *Situation* container holds several tracked variables that describe the current interaction context. The variables shown in this excerpt are

- "own_contribution_need": the intrinsic urgency of the communicative act that the agent wants to perform

- "other_voice_state": the observation that the other participant is being silent right now

- "own_phrase_state": the current execution phase of the agent's pending contribution

- "delay": the time that passed since the agent's last attempt at speaking

The *Communication Memory*, in contrast, holds the messages that the agent observed. In this example, the agent remembers that the user's gaze shifted to the side, that the user started speaking, and that the agent successfully offered to sell the user a vacuum cleaner.

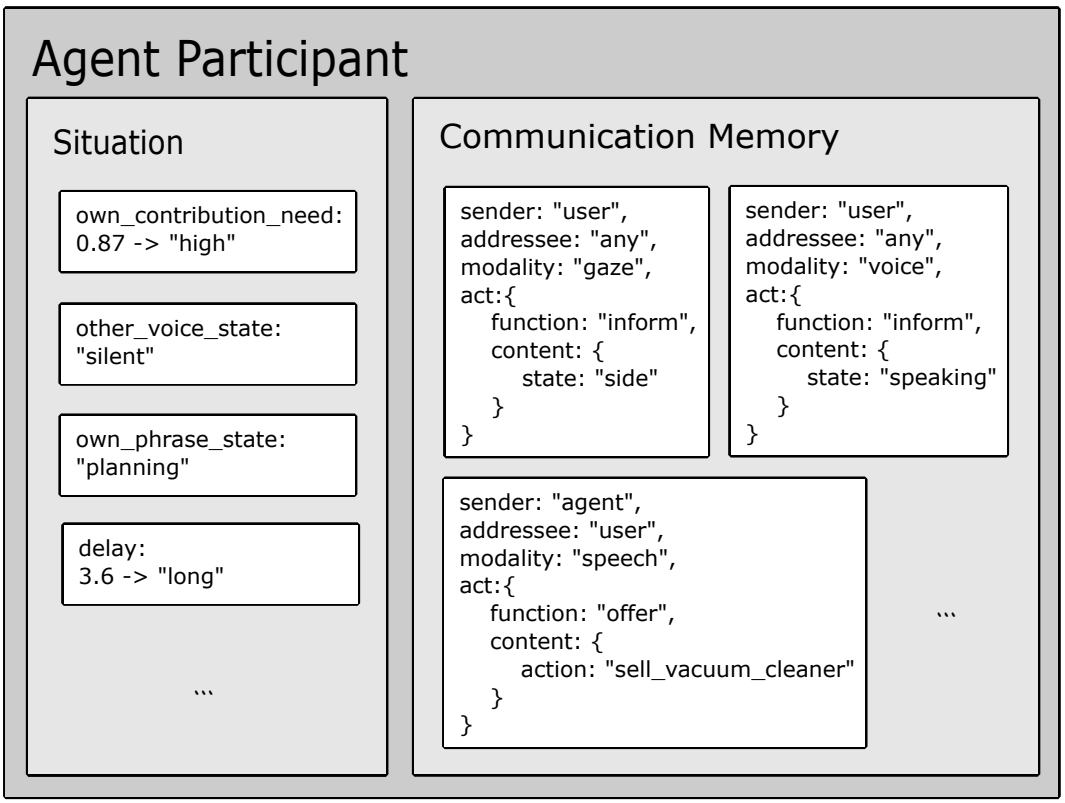


Figure 7.1: Information storage for an agent participant, holding monitored situation parameters and exchanged messages. *Left:* Example variables that represent the urgency of the pending utterance, the last known state of the partner's voice activity, the agent's progress in delivering the current utterance, and the time elapsed since the agent's last attempt at speaking. *Right:* Example messages sent by both participants. Shown are two input events from the user, specifically a gaze shift to the side and the start of voice activity, as well as the agent's spoken offer to sell the user a vacuum cleaner.

This section explains how this knowledge is defined.

7.2.1 Exchanged Messages

In this thesis, exchanged information is represented in a manner similar to the DiAML standard by Bunt et al. [24] (see also section 2.4.2). These communicative acts will be used for both human and artificial participants of the interaction, making it easier to simulate conversations between two agents first

and replace one of them with a human user later. A detailed description of these acts can be found in section B.1 of the appendix.

As explained in section 5.2.1, an interlocutor needs a minimal amount of information before they can start responding to what the other says or does. In other words, they need to know what the communicative function of the message will be, and possibly additional content that they can use to plan their response.

The semantic interpretation of a specific message is highly dependent on context, especially when it comes to nonverbal signals. Therefore, recognizing a particular phrase or gaze direction is not enough to establish the common ground (see section 3.4). Instead, the memory of the agent only holds the information that can be extracted directly while the rest of the semantic interpretation is left to the dialogue manager.

Speech

The earliest plausible time for an agent's response is the point at which a human would be able to deduce the communicative act. For an artificial agent whose responses are scripted, this is rather straightforward to implement. The main requirement is that the agent's behavior realizer provides feedback on the execution of commands.

However, this is impossible in an interactive setup with natural human speech. In that case, an incremental speech parser needs to be used. The MNI can be considered complete when said parser is able to provide a communicative function and the associated content that the dialogue manager requires for proceeding. Depending on the surface form of the message and the complexity of the domain, this can happen anywhere between the first syllable and the end of the utterance.

For example, the utterance "no thank you, I'm not interested" can be parsed as a rejection right after the "no". In contrast, the sentence "I would like to present our newest vacuum cleaner to you" would only make sense after the addressee has heard the words "vacuum cleaner". (Assuming that the message is purely verbal and the vacuum cleaner was not mentioned before.) While "I would like to present" already reveals the communicative function as an *offer* for information, the addressee does not yet know what kind of information is offered. Consequently, the dialogue manager would ignore the early NLU result that does not contain a value for the content and wait for a later result that can provide the missing information.

In other words, the incremental speech recognizer may provide any number of intermediate parsing results, but only the dialogue manager knows when enough information was received.

Other Modalities

Compared to speech, the modalities of voice activity and gaze direction carry little intrinsic meaning. Therefore, they need to be interpreted by the interaction manager that has access to other context information.

For example, silence can occur during both attentive listening and pauses within a speaker's turn [121]. Likewise, gaze aversion directions are not exclusively linked to one cognitive state, as seen in the probability distributions found by Andrist et al. [4]. (Refer back to section 5.3 and figure 5.1 for details.)

Therefore, signals in these modalities are interpreted as communicative acts with the function of *informing* the observer about a new *state*. Said state is given as an attribute of the act's content.

The voice activity can take on the states "silent" and "speaking". Gaze states can be a basic direction label, such as "left" or "up", or any named target in this interaction domain. For example, in the scenarios that were implemented for this thesis, gazing directly into the agent's camera is mapped to the state label "partner".

Updating the influence diagram (see chapter 6) with these observations allows for reasoning about the other participant's cognitive state. For example, the knowledge that the user is looking at the agent may imply that they want to receive more information from the agent.

7.2.2 Situation Parameters

The current interaction context is represented by several variables called *situation parameters*. Each one has a unique *name* and a current *value* that is either numeric or categorical. In this thesis, specifically the interactive prototype from chapter 10, situation parameters are used in different ways.

Intermediate Storage

To keep track of gaze shifts or voice activity, the agent stores the most recently observed states as categorical parameters. Different parameters are used for the agent's own signals and those perceived via its sensors. For example, a parameter named "own_gaze_target" would be set whenever the agent's animation component turns its head or eyes. In contrast, one called "other_gaze_target" would copy the "state" value from the most recent communicative act that was perceived via the "gaze" modality.

Tracking Durations

A parameter value can be updated periodically to track the elapsed time since a triggering event. For example, the value of a parameter named "delay" can be increased by the number of milliseconds passing between each automatic update. Once the agent attempts to speak, this value can be reset to zero.

Simulating Needs

A parameter value can also be read from a table, for example, to map the communicative function of the planned utterance to its urgency or that of the recently finished one to the need for a response. Both of those mappings are implemented in the interactive prototype for this thesis.

7.3 Information Exchange

Artificial agents need ways to "see" and "hear" both other agents and human interlocutors. Consequently, semantic information needs to be explicitly transmitted between all kinds of participants.

7.3.1 Sending and Receiving

An agent can only respond to a message if it can access the included communicative acts. Therefore, the sender needs to store those where the addressee can retrieve that information. For example, a central knowledge base could hold all messages, or a central hub could distribute copies of specific messages to every addressed or overhearing participant.

Ideally, the same infrastructure can be used for both artificial and human participants. It should also be independent of the modalities and the interaction domain so that it can be reused for a wide range of scenarios. The communicative acts explained earlier facilitate the design of such a general infrastructure. In this thesis, all exchanged messages will be mapped to these acts, regardless of who produced them in which modality.

7.3.2 Memory Retrieval

Messages are stored as feature structures similar to those described by Mehlmann et al. [83, 84]. Each of them contains the following attributes.

- **Time:** The moment when the message became available to the participants.

- **Sender:** The name of the participant who produced this message.
- **Addressee:** The name of the participant who is supposed to receive this message.
- **Modality:** The modality over which this message is transmitted.
- **Act:** The communicative act transmitted by this message.

As explained before, the communicative act contains the associated function and any content necessary for acting on it. The content, in particular, is a collection of attributes that can be atomic or be themselves collections of attributes.

The stored messages are used to advance the dialogue by selecting a response and deciding whether an agent needs to retry a certain contribution. Any of the attributes in a message can be used to query a participant's memory. The goal is to support simple conditions, such as the presence of voice activity, or complex ones, like whether the user asked a specific question.

7.4 Synchronization Between Components

This thesis aims to create a general turn-taking model that works in different interaction setups. For example, the initial tests involve conversations between two computer-controlled characters, whereas a later setup has a human talking to a social robot. Another goal is to combine machine learning approaches such as reinforcement learning with this turn-taking model, tailoring the agent's personality to the user's requirements while ensuring that the behavior is consistently derived from those traits.

Consequently, the decision-theoretic model has to be embedded in a suitable architecture that provides it with the necessary information and passes the decision on to the dialogue manager, regardless of who or what the participants are. The different forms of knowledge and reasoning results must be communicated between the different components of the dialogue setup. The appropriate observations need to be set in the influence diagram, and the dialogue manager needs to act on its decisions.

Figure 7.2 gives an overview of the information flow within the proposed architecture. Semantic information is used to advance the dialogue flow. In contrast, surface observations, such as voice activity or execution progress, are tracked as context parameters and communicated to the influence diagram at appropriate moments. Within the dialogue manager, there is a further separation between preparing a dialogue contribution and actually sending

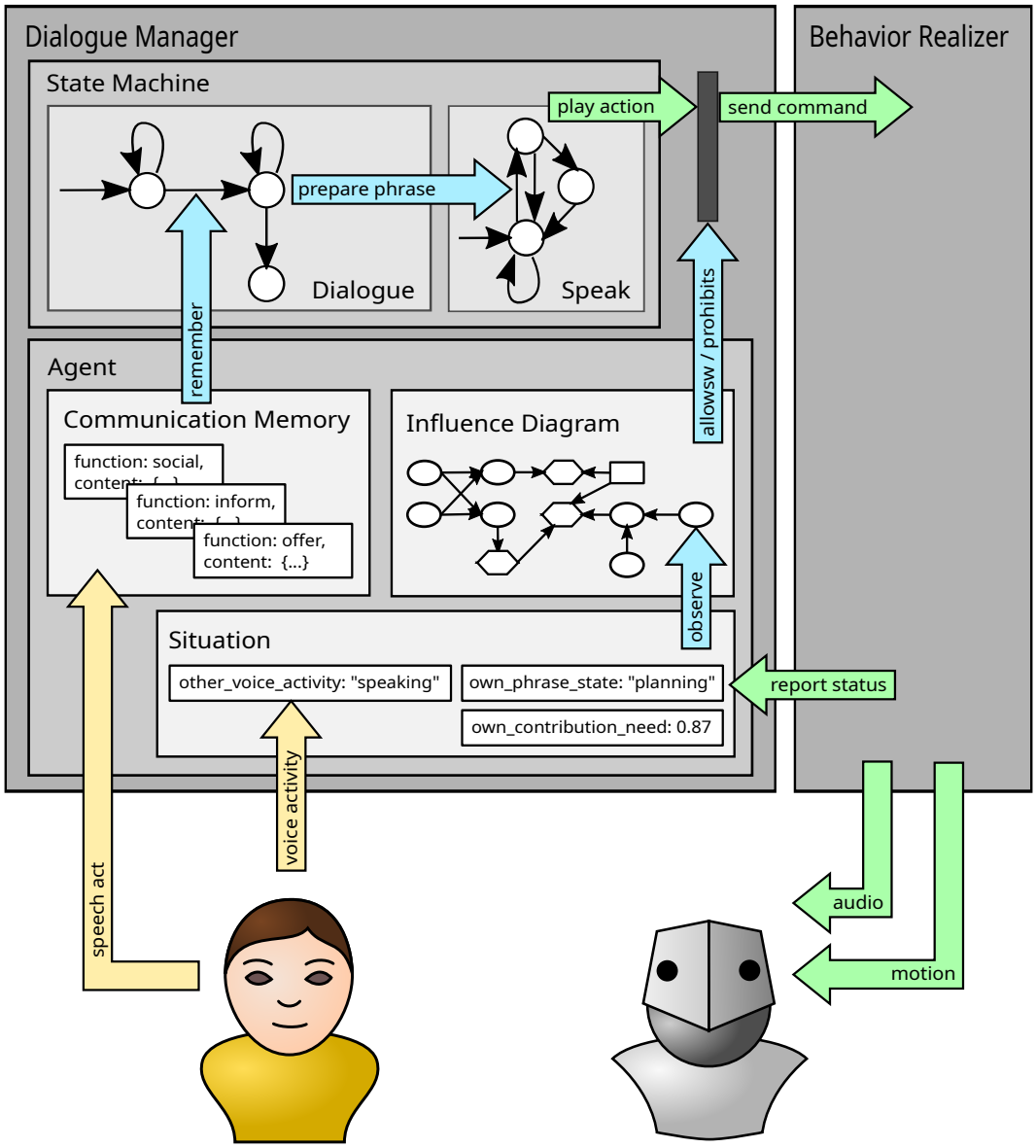


Figure 7.2: Flow of information between the different components of the dialogue setup.

the command to the agent for execution. At this point, the influence diagram is able to stall or permit the sending of the command.

7.4.1 Behavior Definition

The dialogue manager chooses behaviors in connection with the intended meaning, but the agents only process them on the level of their surface appearance. This separation between semantics and form makes it easier to implement different interaction scenarios. Furthermore, the technical realiza-

tion of the behavior is separated from the turn-taking logic of the agent so that it can be used with different graphical or robotic characters. More information about behavior realization will follow in chapter 8.

Verbal Behavior

On the agent's side, the communicative act is anchored to a bookmark in the text and is handled as soon as this checkpoint is passed during output. While the words that an agent should speak are sent to its behavior realizer, only this semantic representation is committed to their memory. Likewise, only the communicative act extracted by the NLU is forwarded to the agent, but not the actual words that the user said.

Gaze Behavior

Gaze targets are defined as labeled sets of coordinates. The dialogue manager uses these labels to activate a particular target, at which point the label is mapped to the equivalent coordinates for this agent. From then on, the agent interpolates between the coordinates without considering the labels.

However, the interpolated result is later discretized again so that a message can be sent to other participants. Both the agent's and the user's gaze direction are mapped to the closest labeled target.

7.4.2 Tracking Situation Parameters

Before they can be fed into the decision-theoretic model, the different context factors need to be monitored and discretized. Therefore, each agent has its own collection of *situation parameters*. As described in section 7.2.2, some are updated when the agent receives new messages from other participants, for example, when the voice activity stops or when the interlocutor's gaze shifts. Others, such as the agent's speech desire or the time elapsed since the last contribution, change over time or upon certain trigger events in the behavior scheduling.

To avoid slowing down the turn-taking decisions, the influence diagram is not updated automatically when one situation parameter changes. Instead, the values of the situation parameters are explicitly transferred to the influence diagram at those points when a change in the decision is likely. Specifically, this happens when the agent proceeds from one phrase state to the next, when its behavior realizer reports a change in execution status, or when a message is exchanged between the participants. The expected utility for the available actions is only calculated after all observations have been set to reflect the current situation.

7.4.3 Regulating the Dialogue Flow

The core idea is that the agent's influence diagram acts as a gatekeeper between the dialogue manager and the behavior realizer. While the dialogue manager is responsible for providing the next contribution, the influence diagram determines the precise moment when speech is started or stopped. This approach allows for a distinction between the moment when the agent could theoretically speak and the one when its personality dictates that it should.

Whenever an utterance becomes available to the agent, it needs to wait until its verbal attention shifts from the other participant to the agent itself. While executing the speech command, the agent continues to monitor the influence diagram. The speech command is immediately canceled if the verbal attention shifts back to the other party.

Depending on the progress before the cancellation, the agent may retry its contribution or move on to the next one. Specifically, the agent queries its semantic memory to see if the MNI has been transmitted already. The dialogue advances only when the agent remembers speaking that information.

7.5 Extensions

The participant types in this framework are meant to be extended with different capabilities. For example, users are handled differently than scripted agents or those agents controlled by an influence diagram.

7.5.1 Common Elements

One thing that all participants have in common is the connection to a central message hub. Every form of communication, from the user's inputs to the current state of the agent's gaze and the spoken sentences of both, is sent there so that all artificial participants can perceive them.

In addition to that, artificial participants have their own local storage for situation parameters and messages that they received.

7.5.2 Interruptible Participants

The feature that distinguishes interruptible agent participants from basic agents is the influence diagram that they use for regulating the dialogue flow. In the current version of this framework, further additions include a gaze animator or the connections to the dialogue manager and the agent's embodiment.

7.5.3 Learning Participants

Some initial tests have been done with adding machine learning capabilities for online personality adaptation. While they are outside the scope of this thesis, they did influence the development of the Participant Framework.

Most importantly, the tracking of situation parameters was implemented with the creation of state labels in mind. An additional property of each parameter determines whether it is included in the state definition, and their discretized values are then combined to uniquely identify a world state for approaches such as reinforcement learning.

An extension of the Interruptible Participant could, for example, change the personality configuration at runtime by setting the observations at the influence diagram's trait nodes.

7.6 Conclusion

This chapter presented the architectural framework in which the developed turn-taking model was embedded. A suitable infrastructure was developed for allowing users to communicate with different types of agents or allowing such agents to communicate with each other as they would with a human.

Regardless of the sender or the modality of a message, its meaning is represented as a communicative act inspired by the DiAML standard (see section 2.4.2). This semantic information is used to advance the dialogue, while contextual information is used to infer the most beneficial actions via decision-theoretic reasoning. For this, the influence diagram is updated with the tracked situation parameters as the agent progresses through its dialogue contribution. This separation between the context information and the turn-taking model facilitates the tracking of continuous values or the simulation of the agent's needs, all of which are only discretized when needed.

The decision of the influence diagram, in particular the verbal attention target, then regulates the information flow between the dialogue manager and the agent's behavior realizer. Speech commands are only sent after the agent's attention shifts toward itself, and they are canceled as soon as it shifts to the interaction partner, putting the agent into listening mode.

The realization of the behaviors will be explained in the next chapter.

Chapter 8

The RobotEngine Framework

8.1 Introduction

Nowadays, a wide range of graphical and robotic agents is available to researchers, and the number of options continues to increase. Consequently, there have been approaches for unifying the software environments and separating the reusable interaction logic from the agent-specific execution details. (See section 4.4 for more information.)

During this thesis, several very different agents were explored for use with the turn-taking model and testing its generated behavior. This called for an easy way to swap out one agent type for another or to migrate the existing application logic to a new model. Therefore, the RobotEngine framework was developed as an interface between the main application and various agent types. It was built around the following requirements.

- **Modular Setups:** The main goal is to flexibly combine different graphical and robotic agents with different control applications to speed up the development of prototypes. To do so, as little as possible should have to be reimplemented for a specific agent. The interaction logic, specifically, needs to remain general enough that it can work with any of the supported agents.

- **Hiding Implementation Details:** To support this modularity and make prototyping easier, the control application should require as little information as possible about the agent's technology. For instance, it should not need to know the manufacturer's naming conventions or how conflicting commands are scheduled internally.

- **Human-readable Communication:** Since many different components need to be connected, the setup can quickly become confusing and hard to debug. Therefore, the messages exchanged between those components should be straightforward to read or manually create during testing.

This chapter gives an overview of the technical requirements for having an agent display the proposed turn-taking behavior. After that, it presents the design principles of the RobotEngine framework and how they address said requirements. Finally, it looks at several agents used during this thesis and the functionality that the framework makes accessible for those. A summary concludes the chapter.

8.2 Technical Requirements for Turn-Taking

While comparing different agents, it became apparent which platforms were suitable for testing the turn-taking model and why. Some platforms could be extended to meet the requirements, for example, by adding explicit scheduling for concurrent animations. Other issues, such as the API not exposing all functionality of the TTS engine, could not be solved. This section sums up the most important properties that an agent platform needs to have so that it actually benefits from the proposed turn-taking model.

8.2.1 Flexibility of Output

Human communicative behaviors are highly flexible. Signals change from one moment to the next, depending on the reaction of the other person. If these patterns are to be transferred to artificial agents, similar flexibility is a fundamental prerequisite. In particular, the agent's software must provide detailed control over the execution of audio and animation commands.

Asynchronous Behaviors

Turn-taking signals such as gaze are interleaved with the turn on a very fine time scale. They must also be adjusted dynamically based on what the other party is doing. For example, if one participant tries to interrupt the other, the active speaker may need to respond by turning their face away.

Therefore, it is essential that audio output and servo animations can be controlled independently and in parallel. Preferably, this concurrency should also be supported for other modalities, such as LED color changes or the individual axes of a robot's neck.

Feedback regarding Execution Progress

From the need for parallel, non-blocking command execution follows the need for ways to track their progress. For example, the application triggering a speech command must be informed when the audio output starts, when it reaches the end of the MNI, and when the output stops. A gaze animation process should know the current head and eye direction to adjust the interpolation speed and avoid overly abrupt motions.

Consequently, the agent's software needs to raise feedback events or execute callback methods upon reaching certain checkpoints. Otherwise, the turn-taking model will not be able to reason about the agent's "cognitive" processing load or track the amount of time that the agent has been looking at the user.

Canceling of Started Commands

When the goal is to model interruptions and turn yielding, the agent's behaviors obviously need to be interruptible in the first place. In particular, the agent's speech output must provide a canceling command.

Furthermore, procedural animation is preferable to keyframe sequences because the latter usually need to be planned in advance. If keyframe animation should be necessary, small building blocks are recommended. For example, rotating the neck for less than a second is preferable to animating a full sequence of looking to the side, fixating an object there for some time, and returning to a neutral gaze direction.

A related topic is the buffering of commands in case of resource conflicts. If the dialogue context should require a change of plans, the agent's software must allow for the cancellation of pending commands before they are dispatched.

8.2.2 Modularity

Modularity is a common design choice in software development because it allows for reusing existing code, thereby saving time and resources. For developing a model of communicative behavior, it becomes even more important due to the wide range of possible interaction domains.

Separated from Interaction Domain

The execution of commands should not require semantic information so that it can be reused for different application scenarios.

Furthermore, by separating the surface behavior patterns from the communicative intention, it becomes possible to model the fact that the same message can carry different meanings in different situations. It also keeps the action

space smaller because there is no need to prepare all combinations of behavior and meaning.

Separated from Agent Technology

Authoring the behavior should not require knowledge of how the agent's functionality is implemented. Parameters such as the agent's personality traits or content such as the spoken text need to be accessible to domain experts who may not have a technical background.

A high abstraction level also makes it easier to swap out one agent for another. For example, if gaze behavior is defined by a set of coordinates, it does not matter whether the animation is executed by panning a 2D image of graphical eyes or by rotating the physical eyeballs of a robot. If two different robots have similar TTS capabilities, such as bookmark events and a canceling command, the speech timing implemented for the first one can be transferred to the other with little effort.

8.3 Design Principles of the Framework

The RobotEngine Framework aims to bridge the gap between the behavior model developed for this thesis and the various agent platforms that are controlled differently. This is done via a standardized messaging protocol that hides the agent's implementation details from the application controling its behavior.

While it was developed around the Visual SceneMaker, it takes inspiration from the SAIBA framework and especially the ASAP realizer [73]. For example, it builds on the clear separation between planning and executing the behavior, with semantic content only playing a role on the dialogue manager's side. At the same time, it requires the behavior realizer to report back on the execution progress, as proposed by Kopp et al. [73].

This section takes a closer look at the components of the RobotEngine framework and the information exchanged between them.

8.3.1 Messaging Protocol

The RobotEngine Framework primarily defines a messaging protocol that needs to be supported by all involved agents and control applications. Said protocol decouples the agent software from the implementation of the dialogue logic and provides a uniform interface between the different components.

Message Format

The messages exchanged between the agent and the control application are designed to be human-readable with a standardized but flexible text representation. They describe the observable surface behavior without attached semantics. This abstraction level was chosen so that semantic mappings would not have to be re-implemented for different agent platforms.

- **Command Message:** A command consists of a *task identifier*, an action *type* referring to the modality, and a list of *parameters* that are needed for executing the command. For example, this could be the text to be spoken, the name of an animation file, or a set of coordinates for the gaze direction.
- **Status Message:** A status update references the associated command via the *task identifier* and holds information about the execution progress. At the very least, the *status* property identifies the reached checkpoint, such as the start or the end of the command execution. Optional *details* provide additional information. For example, this could be the identifier of a bookmark reached in the speech output or a reason why a command cannot be executed.

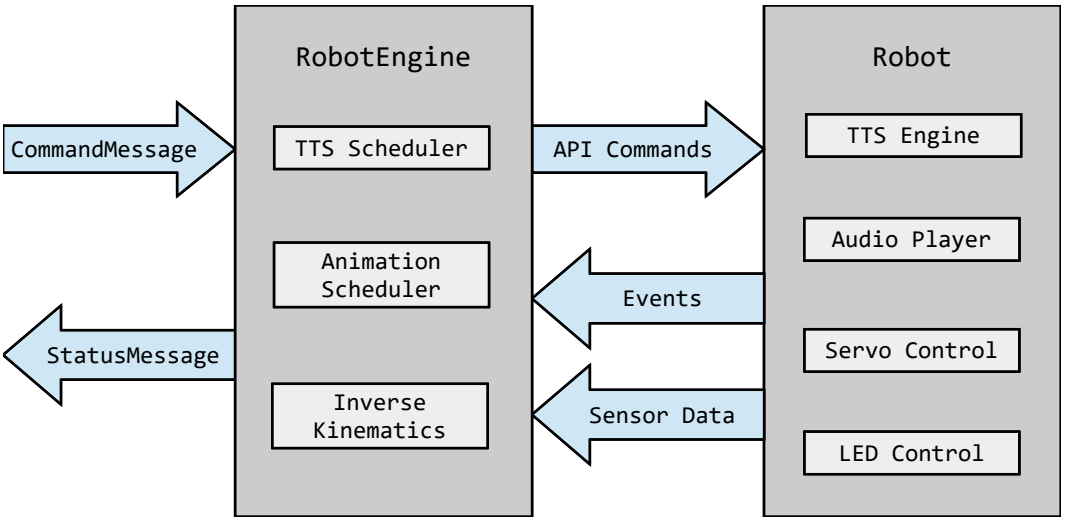


Figure 8.1: Translation between the agent's API and the standardized messages via the RobotEngine component.

Information Flow

Two major components are required. One is the *control application*, which can be a Wizard-of-Oz experiment remote control interface, a simple script

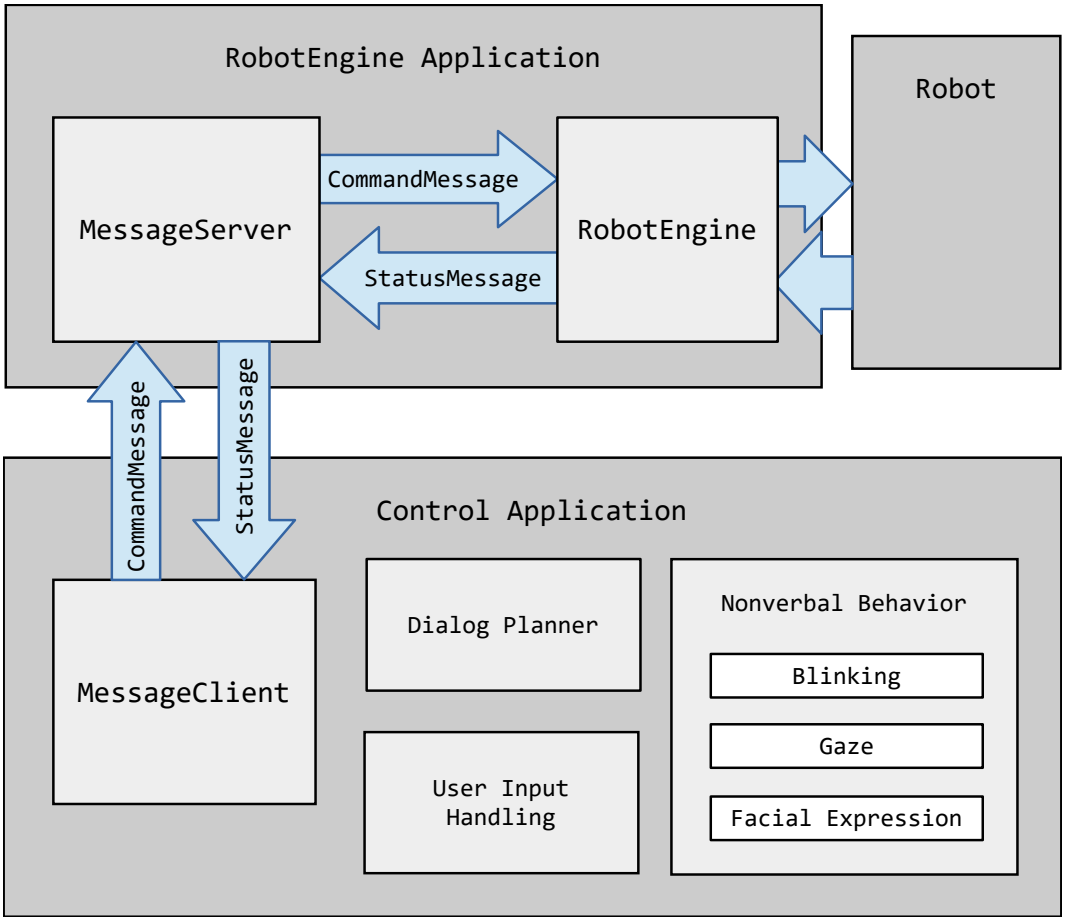


Figure 8.2: Information flow between the components within the RobotEngine framework.

player, a complex state machine, or any other approach for choosing the agent's behavior. It translates these behaviors to command messages, sends them to the agent, and uses the associated status messages to decide how to proceed. For example, it might wait for the status "finished" before moving to the next line of the script or trigger an action that is tied to a bookmark in the speech command.

On the other end, there is a *RobotEngine* implementation that was tailored to the specific agent platform. It translates the command messages to the necessary API calls for the agent's software and, if necessary, converts standardized parameters to those required for this particular platform. It then makes the agent perform the requested action, uses whatever method is available to monitor the execution progress, and sends appropriate status messages back to the control application.

8.3.2 Modularity

To further support the reuse of code, several components need to be incorporated in the RobotEngine and control application implementations. The goal is to minimize the amount of work necessary for connecting a new agent platform or control application to the existing setup.

Core Components

The following core classes have already been implemented in Java, and to some extent in other programming languages[1].

- **RobotEngine:** An abstract base class that supports the message protocol and is meant to be extended with agent-specific methods for executing the commands.
- **StatusMessageHandler:** An interface that has to be implemented by the control application to process the status messages.
- **Messaging Classes:** These comprise the *CommandMessage* and the *StatusMessage*, as well as a *MessageServer* and *MessageClient* for connecting to external applications via UDP.

Agent-Specific Components

Graphical and robotic agent platforms often use proprietary libraries, which makes it difficult to distribute specific implementations of the RobotEngine class. Therefore, certain classes will need to be re-implemented for whatever agent one wants to use.

- Incoming **command messages** need to be mapped to method calls in the agent's API.
- Their **parameters**, such as the names of the servo to move or the unit of its target position, need to be mapped to their equivalent on this particular platform.
- Low-level **scheduling and conflict handling**, for example, of requested speech commands or parallel animation tasks, needs to be implemented based on the callback events or status variables provided by the agent's API.
- Available **feedback** needs to be translated to equivalent status messages that are sent back to the control application.

[1] They are available at https://github.com/kjanowski/RobotEngine

8.4 Supported Agents

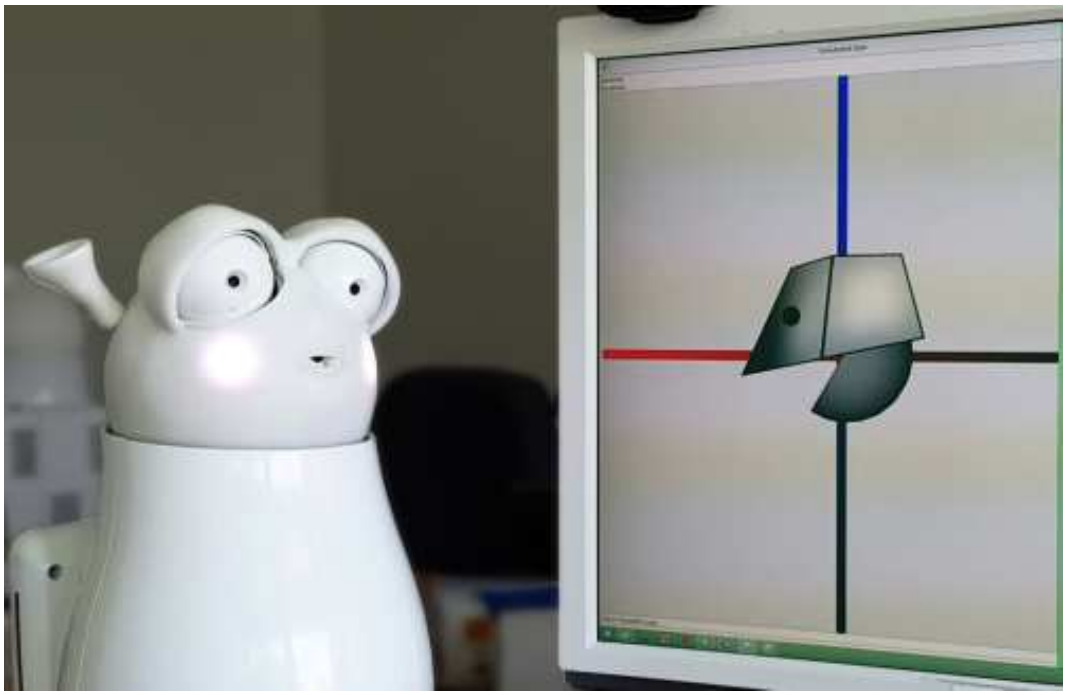


Figure 8.3: A Robopec Reeti robot in conversation with a graphical Klappmaul agent.

Several robotic and graphical agents were considered for testing the turn-taking model that was developed in this thesis. The RobotEngine framework made it easier to exchange one agent type for another or to combine different agents in the same application. This section gives an overview of the most relevant agents for the presented research.

8.4.1 Aldebaran NAO

The NAO robot was developed by the French company Aldebaran[2]. It was mainly used in the early stages of this thesis for exploring gaze behaviors related to grounding [85, 83, 84].

Two model versions were available, the V4 and V5. Both were controlled through the same Python API, so only one "NaoEngine" was used for both. It evolved from a software "proxy" that had been developed for the early experiments and served as one of the reference implementations for the entire RobotEngine framework.

[2] https://www.aldebaran.com/en

Animation

The NAO is a humanoid robot with a total of 25 DOF. There are two in the neck (pitch and yaw), five in each limb, one in the pelvis, and one in each hand.

It can play back keyframe animations that have been created with Aldebaran's software. Since that software only exported them as standalone Python scripts, an additional converter was implemented to extract the raw animation data and store it in a more generic format.

Additionally, the servo motors can be animated dynamically through the API. The source code for procedural gaze and pointing animations already existed thanks to earlier projects at the chair [59].

Text-To-Speech Output

The NAO uses Acapela for TTS output. It comes with English and a selected second language installed. At the University of Augsburg, German was chosen as the second option.

Cancellation of pending speech jobs is supported by the Python API, so it was straightforward to implement the stopping command. Furthermore, the volume is directly adjustable, but the pitch can only be set indirectly via a "pitch shift" parameter that changes the overall timbre.

8.4.2 RoboKind R-50

The RoboKind R-50 was developed by Hanson Robotics[3]. During this thesis, the model was discontinued and also turned out to be unsuitable for the presented turn-taking research. Nevertheless, the efforts made to connect it to various control interfaces provided the groundwork for the RobotEngine implementation as a whole.

Animation

The most prominent feature of the RoboKind R-50 is its articulated face. There is one DOF for the inner eyebrows, one for the eyelids, one for the eyes' pitch, two for each eye's yaw, one in each lip corner, and one for the jaw. Figure 8.4shows several example expressions.

Apart from the face, there are three DOF in the neck and several more in the humanoid body. Additional scheduling inside the RobotEngine resolves conflicts between contradictory animation commands while ensuring that different servos can move in parallel.

[3] https://www.hansonrobotics.com/

Figure 8.4: Facial expressions of the RoboKind R-50 Zeno. From top left to bottom right: Neutral, surprise, fear, anger, happiness, sadness, contempt, shame.

Like the NAO, the R-50 was used in the early exploration of gaze patterns, but experience showed that it landed in the Uncanny Valley for many students and colleagues. Therefore, later research used less human-looking agents instead.

Text-To-Speech Output

The RoboKind R-50 uses the same TTS software as the NAO. However, this robot's API does not support the interruption of running speech jobs. Consequently, it was not considered for further use in turn-taking research.

8.4.3 Robopec Reeti

The Reeti robot was one of the main agents used during this thesis. It was developed by the French company Robopec[4].

The V1 and V2 models offered similar functionality but were controlled differently below the surface. While the V1 was based on the Urbi framework, the V2 was based on ROS. Both were programmed through very different Java APIs, so the RobotEngine framework played an important role in using them interchangeably. Two versions of the RobotEngine were implemented, one for each model.

[4] https://www.robopec.com/en/constructions/others/reeti-robopec/

Animation

One advantage of the Reeti robot is the cartoon-like face that is very expressive without falling into the Uncanny Valley. It has three DOF in the neck, two in each eye, one in each eyelid, and four in the mouth. Additionally, the ears can turn toward the front or the back. Figure 8.5 shows several examples of facial expressions on the V2 model.

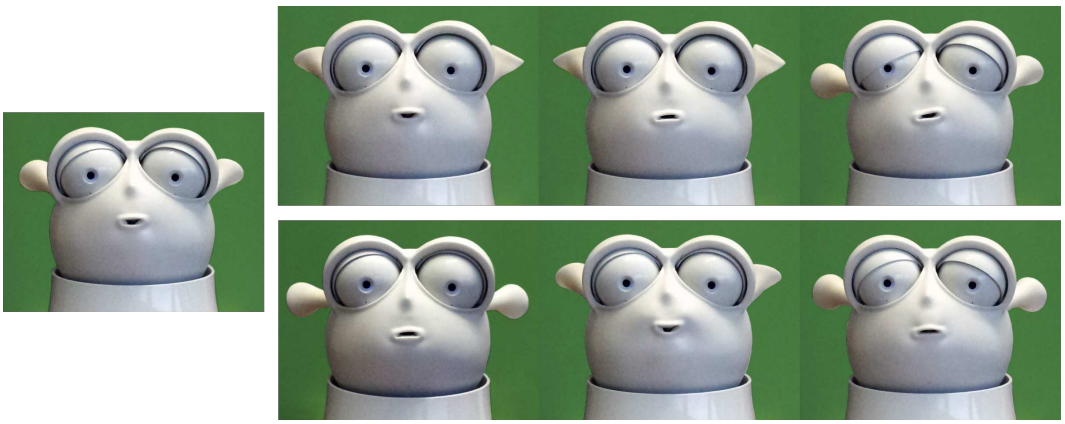


Figure 8.5: Facial expressions of the Reeti V2 robot. *Left:* Neutral expression. *Right:* Different emotional expressions. From the top left to the bottom right, these show surprise, fear, disgust, anger, joy, and sadness.

Both versions support the playback of animation sequences that were created with Robopec's software, as well as the dynamic adjustment of individual servo motors. For gaze animation, their RobotEngine implementations calculate the yaw and pitch of the head from a set of coordinates relative to the neck joint. Additionally, the yaw of the eyes is adjusted based on the distance to the target.

Text-To-Speech Output

Both the V1 and the V2 came with the Loquendo TTS software installed. However, the input format varied slightly regarding the way special characters were escaped.

Bookmarks and the cancellation of speech tasks were supported by both APIs. A wide range of voices was pre-installed, with male and female voices for many languages. Most of their properties, such as volume, pitch, or rate, could be adjusted.

8.4.4 Klappmaul

The "Klappmaul" agent[5] was originally created to test the RobotEngine framework and its messaging protocol. During this thesis, Klappmaul agents were used to prototype the dialogue before moving on to physical robots or in cases when a full laboratory setup was unavailable. At first, it was named "PseudoBot" to indicate that it was a stand-in for various real robot models. The name was later changed because "PseudoBot" appears to be the name of several other projects. This thesis will use both names for consistency with the publications connected to the non-interactive prototype in chapter 9 [63, 64].

Appearance

An additional advantage of the Klappmaul is its simplified geometric appearance. As opposed to an agent with a human-like expressive face, this was expected to make observers focus on the speech timing rather than facial expressions. Therefore, the Klappmaul consists of an abstract polyhedric head with static eyes, an indicated nose, and a prominent hemispherical jaw that makes it easy to spot talking activity.

Figure 8.6 shows the first version of the model with a minimal polygon count and basic textures. Later, the appearance was refined to make it more pleasing to the eye and more presentable in demonstrations of the prototype application. Subtle curving was added to the surfaces, the back of the head was rounded, and the jaw was decorated with a circuit pattern. This version can be seen in figure 8.7. Finally, the third version saw the addition of shoulders to support the display of different gaze directions. It is shown in figure 8.8, with the rendering used by the JavaFX application.

The 3D model was stored in the Collada file format and imported into a JavaFX application. Furthermore, an option was implemented for setting the color of the jaw, allowing for a visual distinction between multiple instances of this agent. This can also be done at runtime, giving the Klappmaul another modality of nonverbal communication similar to setting the LED color of a NAO's eyes or a Reeti's cheeks.

Animation

While the Klappmaul speaks, the jaw mesh is periodically rotated from its rest position to the maximum opening angle and back. No attempt was made to synchronize this movement with the realized phonemes, partially to keep in line with the abstract character design and partially because of the computational

[5] Its implementation can be found at https://github.com/kjanowski/Klappmaul.

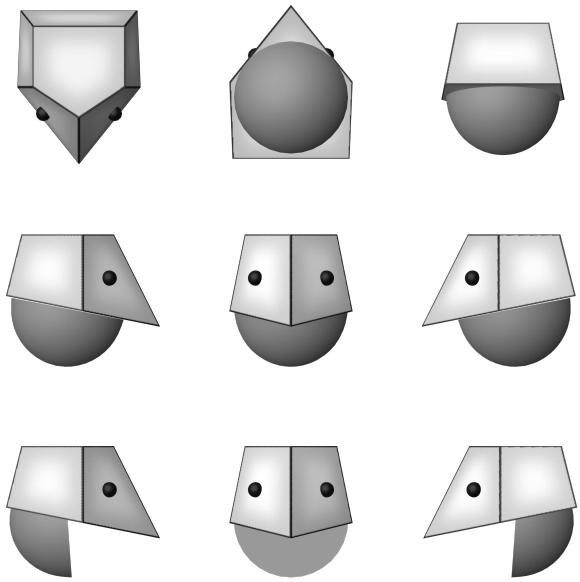


Figure 8.6: The first version of the Klappmaul model.

Figure 8.7: The second version of the Klappmaul model.

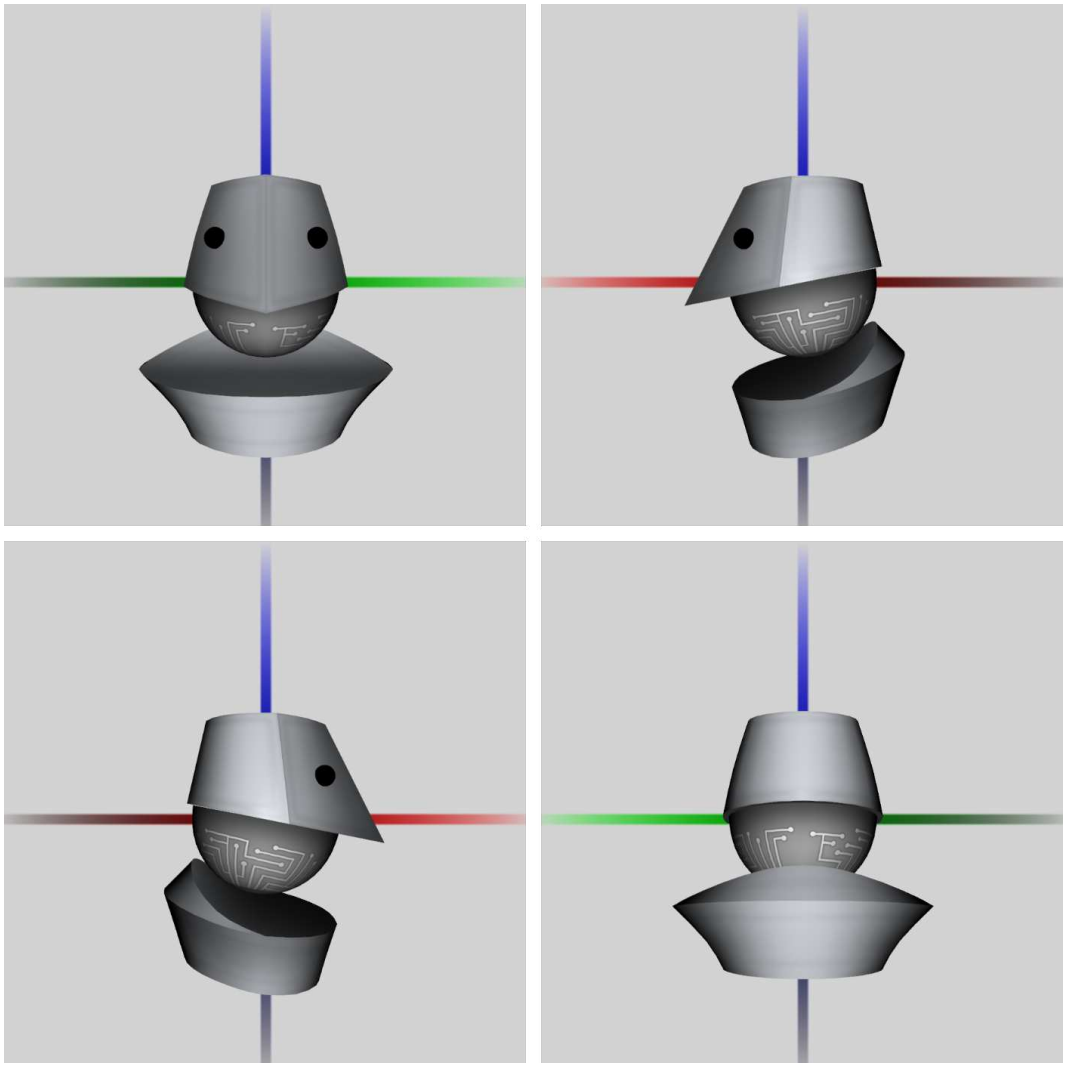


Figure 8.8: The third version of the Klappmaul model, as it is rendered in the JavaFX application.

overhead that would be required for coordinating the animation component with the text-to-speech engine.

The Klappmaul can rotate its head to face in the direction of the given 3D coordinates. Unlike the jaw, the agent's gaze is not animated automatically but only updated when the control application sends an explicit gaze command. The gaze behavior logic was intentionally separated from the agent's behavior realizer to ensure that once a gaze pattern was implemented, it could be easily transferred to any other agent connected to the control application.

Text-To-Speech Output

MaryTTS 5.2[6] was chosen for the Klappmaul's voice. This text-to-speech engine was developed by the German Research Center for Artificial Intelligence (DFKI) and is freely available as open-source software. It supports the Speech Synthesis Markup Language (SSML) and can provide detailed information about the timing of the audio fragments. Furthermore, multiple voices exist for English, German, and several other languages. Most of them can be configured in terms of pitch, volume, and speech rate.

One downside of that TTS engine is that it does not support bookmark events out of the box. Therefore, an additional processing step had to be implemented to extract each bookmark's timing. During audio playback, a second thread has to create the bookmark events at the correct time. This additional processing doubles the time it takes to start a speech command. Therefore, the Klappmaul behavior realizer notifies the control application when the audio output starts so that the dialogue manager can respond appropriately.

8.5 Conclusion

ECA architectures are very heterogeneous. They are based on different conceptual and software frameworks, and there are few standards for social robots or virtual characters. For example, they differ regarding supported programming languages, joint names, parameter units, or value ranges. Furthermore, their software does not always offer the same degree of control or monitoring, which makes it necessary to implement workarounds such as additional scheduling mechanisms.

The RobotEngine framework was developed to provide a uniform interface between different agent platforms and control applications. Decoupling the agent software from the interaction logic facilitates the transfer of the implemented behavior models and test scenarios to different robots or graphical agents. Furthermore, hiding the details of the agent's implementation behind this interface helps focus research efforts on high-level behavior selection.

[6] https://github.com/marytts/marytts

Part III

Proof of Concept

Chapter 9

Agent-Agent Conversation

9.1 Introduction

Developing a turn-taking model is a complex challenge, so it had to be approached step by step. Before confronting it with an actual human, the concept had to be tested in a limited setup. A second computer-controlled agent took the role of the human so that the interlocutor's personality could be varied systematically while the resulting behavior remained reproducible.

The core idea was to simulate the incomplete information that two humans would have about each other, which would then force them to infer each other's intentions from observable behavior. Running in parallel processes with minimal shared information, each agent used its own copy of the influence diagram to adjust the timing of its utterances dynamically.

The influence diagram was designed to represent the connection between a subset of the Big Five personality traits and the simplified goal of *exerting control* over the conversational floor. Among the best-researched aspects in turn-taking are the expressions of *Extraversion* and *Status* (see section 5.3.3 for examples). Interruptions are generally seen as a sign of dominance [113], which has also been confirmed in perception studies with virtual characters [130, 47]. Therefore, Status was a straightforward starting point for building the turn-taking model.

The non-interactive prototype was then realized as a real-time dialogue setup that generated the turn-taking behavior to accompany the playback of a fixed script. The two agents were talking to each other, using the influence diagram's decision to stall, start, or cancel their speech as needed[1].

[1]Examples of this setup can be seen on YouTube at https://www.youtube.com/playlist?list=PLAJ5ZtqkzFRtaO_kK9qPKvjxjzMWawBql

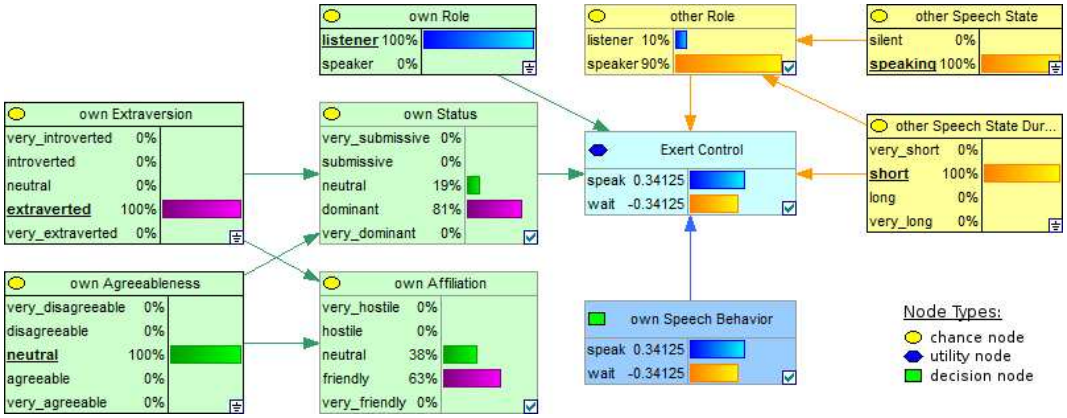


Figure 9.1: Influence diagram used in the non-interactive prototype. The green nodes represent the agent's own configuration and conversational state, the blue nodes represent its reasoning about its goals and available actions, and the yellow nodes represent the agent's belief about its interaction partner. The icons in each node's upper left corner indicate the node type.

To validate this prototype, the agents were configured with varying personalities. Their interactions were recorded as videos. An online survey with 116 participants confirmed that the surface behavior patterns matched the personality traits from which they had been generated.

This chapter will describe the design, implementation, and evaluation of this non-interactive prototype. Significant parts of its content have already been published by Janowski and André in 2018 [63] and 2019 [64].

9.2 Influence Diagram

Figure 9.1 shows the influence diagram at the heart of the non-interactive prototype as it appears in the GeNIe editor[2]. The following sections will explain how its structure was created, how the network parameters were chosen, and how the utilities were calculated.

9.2.1 Diagram Structure

Personality and Interpersonal Attitude

A minimal affect model was chosen for this prototype. On the very left of figure 9.1, there are two chance nodes modeling the personality traits *Extraversion* and *Agreeableness* from the Five Factor Model. These chance nodes are parents

[2]by BayesFusion, LLC, and available free of charge for academic teaching and research use at http://www.bayesfusion.com/

to another pair of nodes that holds the interpersonal attitude dimensions *Status* and *Affiliation*.

Extraversion was chosen because its connection to behavior is more salient and has been more thoroughly researched than that of the other four traits. Furthermore, the relationship between Extraversion, Agreeableness, and the Interpersonal Circumplex has been established in psychological literature. This raised the question of whether this relationship could be accurately reflected with the proposed approach.

Conversation Context

This prototype was developed under the assumption that the interlocutors' behavior mainly depended on their current role in the conversation, as well as that of their partner. Turn-taking conflicts were assumed to occur when both participants see themselves in the same role. For example, when both act as speakers, they fight for control of the conversational floor, whereas awkward pauses slow the interaction down when both consider themselves listeners. Therefore, to reason about potential conflicts, this network models the role in which each participant sees themselves.

In case of the agent controlled by this network, its own role is directly observable. The agent assumes the speaker role when it has content to say and the opportunity to speak. It remains in this role until the important information has been said, either by the agent itself or someone else, at which point it can return to the listener role.

However, the agent cannot directly observe its partner's belief about the latter's role. Instead, it has to rely on surface behavior cues and form a hypothesis of the partner's internal state. To reduce the complexity of this prototype implementation, only speech activity was taken into consideration. Signals transmitted in this channel carry two kinds of information:

- the presence or absence of voice activity
- the duration of the speech and silence phases

The first one alone is insufficient for inferring the other's role since attentive listeners are expected to provide backchannel comments without the intention to interrupt the speaker. Likewise, speakers are briefly silent at phrase boundaries, and there may be disfluencies, such as thought pauses. However, in combination with the duration, the picture becomes clearer.

Here, both signal attributes are modeled as the separate observations "other speech state" and "other speech state duration" to reduce the number of outcomes per chance node. Both are linked to the partner's internal state "other

Role", reflecting their probabilistic interdependence. Furthermore, the duration of speech or silence also indicates the current progress of the partner's contribution or lack thereof. This is an important factor in determining one's own timing, and so the utility for speaking or waiting depends not only on the participant roles but also on said duration.

Conversational Behavior

One well-researched phenomenon in conversation is the connection between speech timing and the expression of high status, also referred to as "dominance" (see section 3.3.1). According to the literature review by Spencer-Oatey [125], high status is often defined in terms of controlling another person's behavior, either with the help of one's strength and resources or due to the authority that comes with a certain role or social rank.

Chulef et al. [29] mention the goal of "having control over others" that is part of the "leadership" cluster of their taxonomy. A similar goal is implied by Brown and Levinson's definition of "negative face wants", a person's desire to act without being hindered by others [20, p. 61-62]. Consequently, the interaction goal "exert control" was chosen for the non-interactive prototype, representing the agent's desire to shape the conversation.

To attain this goal, the agent has two complementary actions at its disposal: It can either *speak* or *wait*. The decision between these two determines the agent's *own speech behavior*, represented by a node of the same name. The goal's utility node models the consequences of these actions. As mentioned above, these depend on the context of the conversation, specifically on the participants' role intentions and the progress of the other party's contribution.

9.2.2 Probability Distributions

Personality

The personality trait variables were discretized into 5 levels each. In addition to the neutral value and the two poles, intermediate levels were inserted to allow for more fine-grained control over the agent's personality. Extraversion had the outcomes "very introverted", "introverted", "neutral", "extraverted", and "very extraverted". Likewise, Agreeableness had the outcomes "very disagreeable", "disagreeable", "neutral", "agreeable", and "very agreeable".

In either case, the prior probability of the five outcomes was uniformly distributed. Although studies exist on the distribution of personality trait expressions, for example, depending on a person's age group [40, 124], these variables will either be known to have a particular value as specified by the

other Speech Duration:	v. short	short	long	very long
other Speech State:	silent			
listener	0.8	0.9	1.0	1.0
speaker	0.2	0.1	0.0	0.0
other Speech State:	speaking			
listener	0.2	0.1	0.0	0.0
speaker	0.8	0.9	1.0	1.0

Table 9.1: Conditional probabilities of the conversational roles given the observed speech behavior.

interaction designer or be derived from the interpersonal attitude configuration instead.

Interpersonal Attitude

Like the personality traits, each interpersonal dimension was represented using five levels. For Affiliation, they were labeled as "very hostile", "hostile", "neutral", "friendly" and "very friendly". The Status levels were labeled as "very submissive", "submissive", "neutral", "dominant" and "very dominant".

As explained in section 3.2.3, there is a deterministic relationship between the Interpersonal Circumplex and the personality traits Extraversion and Agreeableness. Consequently, the mapping between different combinations of these two traits was calculated based on a rotation angle of $-37.5°$. (Refer back to section 6.3.2 for details.)

Conversation Context

As with the personality nodes, the outcome of the "own role" variable - "speaker" or "listener" - would be known from the agent's actual behavior. Therefore, the prior probability was 0.5 for each outcome.

For the conversation partner, a simple heuristic was chosen. The conditional probability of them being in the listener role increased with the observed duration of silence, whereas the probability of them considering themselves the speaker increased with the observed duration of voice activity.

Table 9.1 shows the chosen probability distribution. Since the surface behavior would be observed at runtime, the prior probabilities for their outcomes were uniformly distributed.

9.2.3 Utilities

One commonly used way to define the utilities in a decision-theoretic model is to manually associate each outcome with an explicit cost, for example, on a numeric scale or represented as an amount of money[3]. For example, Fleming and Cohen [42] calculated the weighted sum of the expected duration of the communication, a numeric estimate of how bothered the user is by the interruption, and a numeric estimate of how crucial the decision is. From this formula, they derived the cost of asking the user for a decision with and without the need for additional clarification. Horvitz et al. [56] asked office workers how many dollars they would be willing to pay to avoid receiving specific notifications in different contexts. Similarly, Bohus and Horvitz [17, 18] asked human annotators to rate the gravity of turn-taking errors in video recordings of their system's interaction with a human user.

For this prototype, there were no explicit costs available. However, what was available were various response timings associated with concrete ratings of the apparent personality and interpersonal attitude [130, 26, 47]. Mathematically, choosing between two options only requires the better option to have a higher utility, but the exact distance is irrelevant. Therefore, a "desirability" value for a given timing was used in analogy to the gravity of errors used by Bohus and Horvitz [17, 18].

Identifying Turning Points

One fundamental assumption for defining these utilities was that there were certain turning points at which a character with a given personality and interpersonal attitude would switch from listening to speaking and vice versa. If the other participant's current role were known with certainty, this would lead to deterministic behavior patterns reflecting the ideal timing for that character's configuration.

In the first step, it was necessary to find suitable discretization intervals for the alignment between both participants' speech. To keep the network structure simple, the same intervals were to be used for every combination of speaking and listening behavior. While this increased the number of parents for the utility node, the clear separation between an alignment's type and duration helped structure the utility table and reduce the number of outcome labels per chance node.

Table 9.2 contains the thresholds for which the perceived personality or status were measured. The values reported in the respective sources were normalized to the range of $[-1.0; 1.0]$. Table 9.3 holds additional thresholds that were

[3]Note that Abbas [1] advises against assigning arbitrary values. This reference was not yet available at the time the non-interactive prototype was developed.

Alignment	Aspect	Value	Source
start shortly before the other stops	passive - active	0.80	[130]
	submissive - dominant	0.70	[130]
start after minimal silence	passive - active	0.55	[130]
	submissive - dominant	0.40	[130]
start after 200 ms of silence	submissive - dominant	0.37	[47]
start after 500 ms of silence	introverted - extraverted	0.41	[27]
start after a few seconds of silence	passive - active	0.00	[130]
	submissive - dominant	-0.30	[130]
start after 4000 ms of silence	introverted - extraverted	-0.01	[27]
stop when interrupted during pause	submissive - dominant	-0.04	[47]
stop after minimal overlap	submissive - dominant	0.06	[47]
stop after 1000 ms of overlap	submissive - dominant	0.28	[47]
	introverted - extraverted	-0.01	[27]
don't stop in case of overlap	introverted - extraverted	0.41	[27]

Table 9.2: Timing thresholds that have been examined with regard to the perceived status or personality. Measured aspects have been normalized to range from -1.0 (very submissive) to +1.0 (very dominant).

Alignment	Application Context	Source
backchannel after 200 ms of silence	negotiation training simulation	[137]
restart after 500 ms of silence	neutral quizmaster agent	[18]
start after 600 ms of silence	negotiation training simulation	[137]
start after 3100 ms of silence	adjacency pair response time-out	[108]
start after 3500 ms of silence	neutral quizmaster agent	[18]
stop after minimal overlap	neutral quizmaster agent	[18]

Table 9.3: Timing thresholds that have been used for specific dialogue applications.

not explicitly linked to social dynamics but nevertheless reflect turn-taking patterns that were deemed appropriate for specific dialogue applications. Based on these findings, the following intervals were chosen:

- **very short:** $[0ms; 1000ms[$ (backchannels or phrase boundaries)
- **short:** $[1000ms; 3000ms[$ (short phrases or regular pauses)
- **long:** $[3000ms; 5000ms[$ (long phrases or awkwardly long pauses)
- **very long:** $[5000ms; \infty[$ (overly long phrases or pauses)

Interpolation

Excel spreadsheets were used to systematically combine the participants' roles with the identified timing thresholds. To trigger the behavior switch, the utility for speaking had to be higher if the elapsed time was on one side of the threshold, and lower when it was on the other. The values of -1.0 and +1.0 were chosen for this purpose because the negative number intuitively reflected the idea that the given timing was the opposite of desirable. Therefore, those combinations that were covered by the literature were associated with +1.0 if the character would be speaking at that side of the threshold and -1.0 if they would be silent. Based on this scaffolding, the remaining values were inter- and extrapolated across time and the status dimension. Whenever possible, the interpolation was done linearly, based on the time difference in seconds or the uniformly distributed status levels. To avoid indecision at any given combination, values of 0.0 were manually adjusted to +0.1 or -0.1 based on plausible behavior tendencies for the given status. The values were capped at -5.0 and +5.0, respectively, to keep the utilities regarding *exert control* within a manageable range that would be comparable to future goals.

As a simplification, only the utilities for *speak* were calculated this way. Those for *wait* were defined as the negative value for *speak* under the same circumstances, based on the idea that waiting would have the exact opposite effect. The final utilities are shown in tables 9.4 through 9.7.

duration of partner's silence:	$\leq 1000ms$ very short	$\leq 3000ms$ short	$\leq 5000ms$ long	$> 5000ms$ very long
very submissive	-3.50	-1.50	0.50	2.50
submissive	-2.50	-0.50	1.50	3.50
neutral	-1.00	1.00	3.00	5.00
dominant	1.00	3.00	5.00	5.00
very dominant	3.00	5.00	5.00	5.00

Table 9.4: Utilities for *speaking* when both the agent and the partner are in the *listener* role.

duration of partner's speech:	$\leq 1000ms$ very short	$\leq 3000ms$ short	$\leq 5000ms$ long	$> 5000ms$ very long
very submissive	-1.70	-1.60	-1.40	-1.20
submissive	-1.10	-1.00	-0.80	-0.60
neutral	-0.50	-0.40	-0.20	-0.10
dominant	0.10	0.20	0.40	0.60
very dominant	0.40	0.60	0.80	1.00

Table 9.5: Utilities for *speaking* when the agent is in the *listener* role and the partner is in the *speaker* role.

duration of partner's silence:	$\leq 1000ms$ very short	$\leq 3000ms$ short	$\leq 5000ms$ long	$> 5000ms$ very long
very submissive	1.00	1.00	1.00	1.00
submissive	1.00	1.00	1.00	1.00
neutral	1.00	1.00	1.00	1.00
dominant	1.00	1.00	1.00	1.00
very dominant	1.00	1.00	1.00	1.00

Table 9.6: Utilities for *speaking* when the agent is in the *speaker* role and the partner is in the *listener* role.

duration of partner's speech:	$\leq 1000ms$ very short	$\leq 3000ms$ short	$\leq 5000ms$ long	$> 5000ms$ very long
very submissive	-0.50	-1.00	-2.50	-4.00
submissive	1.00	-0.50	-1.00	-2.50
neutral	3.00	1.00	-0.50	-1.00
dominant	4.00	3.00	1.00	-0.50
very dominant	5.00	4.00	3.00	1.00

Table 9.7: Utilities for *speaking* when both the agent and the partner are in the *speaker* role.

9.3 Implementation

To test this influence diagram, it had to be embedded in a dialogue application. This application needed to advance the script when a character had heard enough, keep the network updated with the various situational variables, and query the network's decision at suitable points in time. To avoid interference from sensor noise or variations in human behavior, the dialogue was to take place between two virtual characters, following the example of ter Maat et al. [130] as well as Glas et al. [47] (see section 5.2). Furthermore, this made it possible to systematically combine different personality configurations for the following perception study.

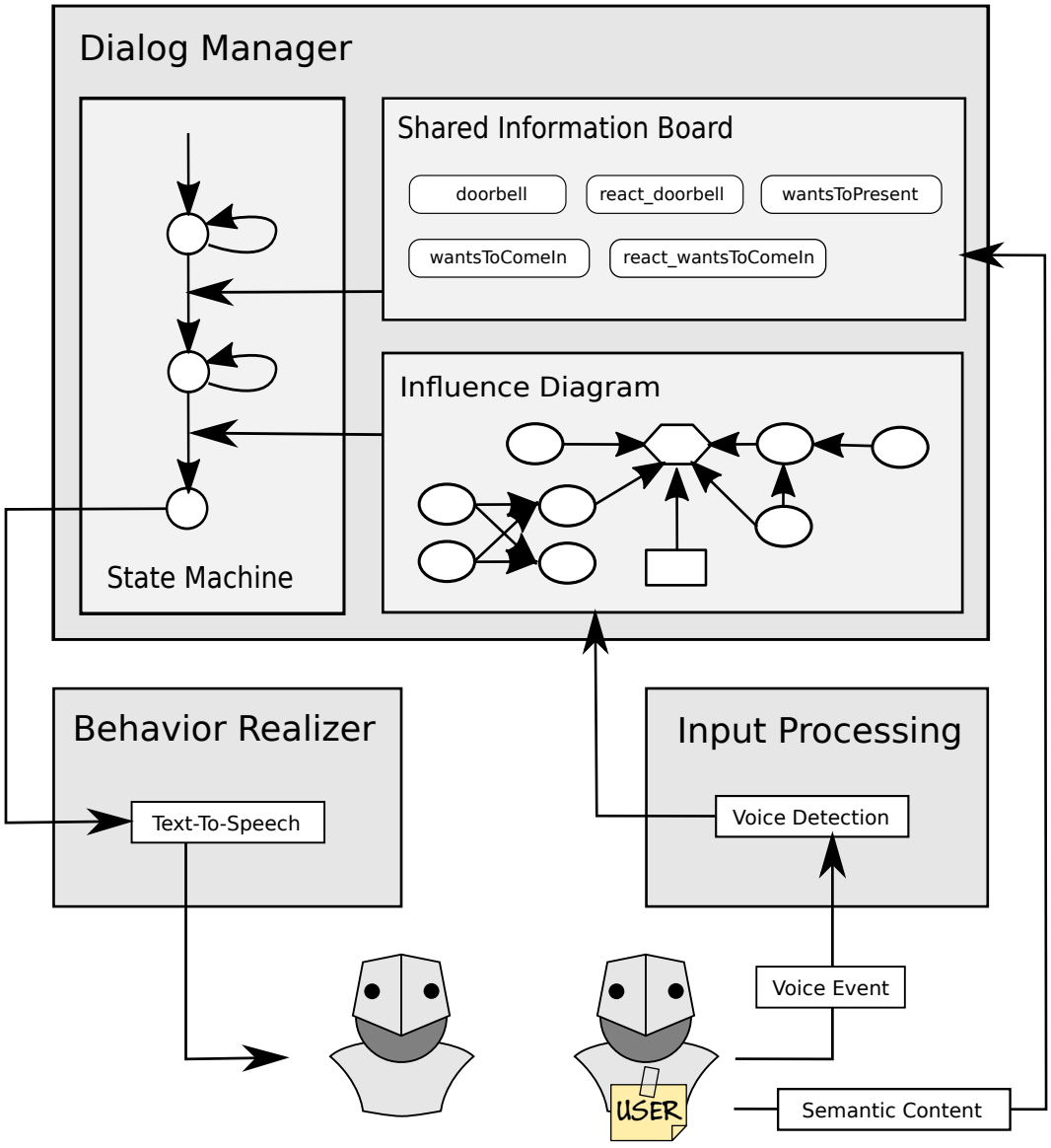


Figure 9.2: Architecture overview of the non-interactive prototype.

9.3.1 Architecture

Figure 9.2 shows the main components of the prototype application. Except for the *Shared Information Board*, all subcomponents exist twice, providing both agents with the same sensing, reasoning, and actuation capabilities. From each agent's perspective, the other agent plays the role of a human user whose internal state can only be inferred from their surface behavior.

Whenever an agent starts or stops speaking, voice activity events are generated within the system and picked up by the other agent's input processing component. The latter's Bayesian network is then updated with the observed speaking or listening behavior.

For the sake of simplicity, the speech act's semantic content is directly forwarded to the *Shared Information Board* rather than parsed from the spoken text. Each participant has a finite state machine that monitors this board and advances the dialogue script whenever the necessary piece of information has arrived. Before sending the speech command to its agent's behavior realizer, the state machine checks the behavior decision of the influence diagram and, if necessary, delays the command until speaking is allowed.

While the agent is speaking, the state machine continues to monitor the influence diagram in case the decision changes. If it does, a stopping command is sent to the agent's behavior realizer, and the state machine prepares to repeat the sentence at the next opportunity.

9.3.2 Dialogue Management

The dialogue manager was based on the Visual SceneMaker[4] that was developed by the DFKI [46]. It models the dialogue flow through hierarchical finite state machines while also serving as the central hub for connecting the different modules. Furthermore, the underlying Prolog fact base is used to store and reason about the speech activity events, as described by Mehlmann et al. [83, 84]. The following plugins were developed for this prototype:

- **InterruptibleExecutor:** Communicates with the PseudoBots[5] via the generic message protocol mentioned earlier. Sends command messages to the respective behavior realizer and listens for status messages about the command execution.

- **SharedInfoExecutor:** A singleton that collects the semantic information exchanged by the agents.

[4] http://scenemaker.dfki.de
[5] The "PseudoBot" was later renamed to "Klappmaul", as explained in section 8.4.4.

- **BayesianNetworkExecutor:** Manages the influence diagram for each agent.

State Machine Structure

On the top level, the state machine consisted of three parts: The *Init* supernode for initializing the dialogue, the *Interaction* supernode for the main phase of the conversation, and a simple node *Wait* providing a short pause before the dialogue restarts. The *Interaction* state is interrupted as soon as both the *Salesperson* and the *Resident* agent have completed their script and set the respective flag.

The *Interaction* supernode holds two separate state machines for the two agents that are run in parallel. Each of those is further split into three distinct parallel processes: The *Contribute* state machine that handles the agent's own speaking activity, the *Dialogue* state machine that keeps track of the agent's script, and the *Observe* state machine that monitors the world state from that agent's perspective.

The *Observe* process consists of one state machine for detecting voice activity events, one for tracking this activity's duration, and one for monitoring the decision node of the agent's influence diagram. These are responsible for synchronizing the sceneflow variables with the events in the Prolog fact base and the state of the agent's Bayesian network.

The *Dialogue* state machine represents the linear dialogue script. It selects the next contribution, while the *Contribute* state machine takes care of executing these contributions based on the world state managed by the *Observe* process.

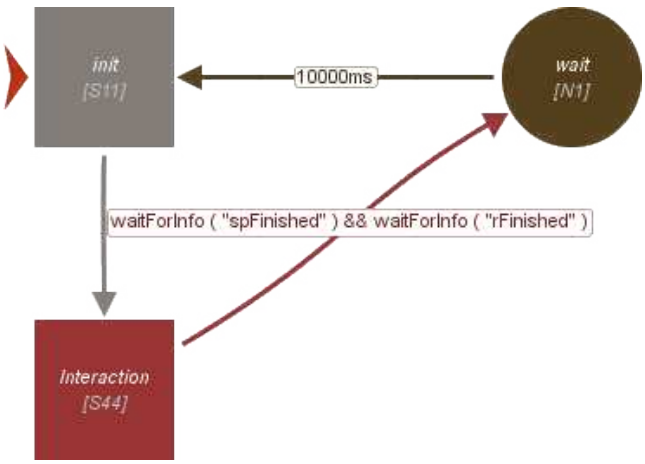


Figure 9.3: The top level of the non-interactive prototype's state machine.

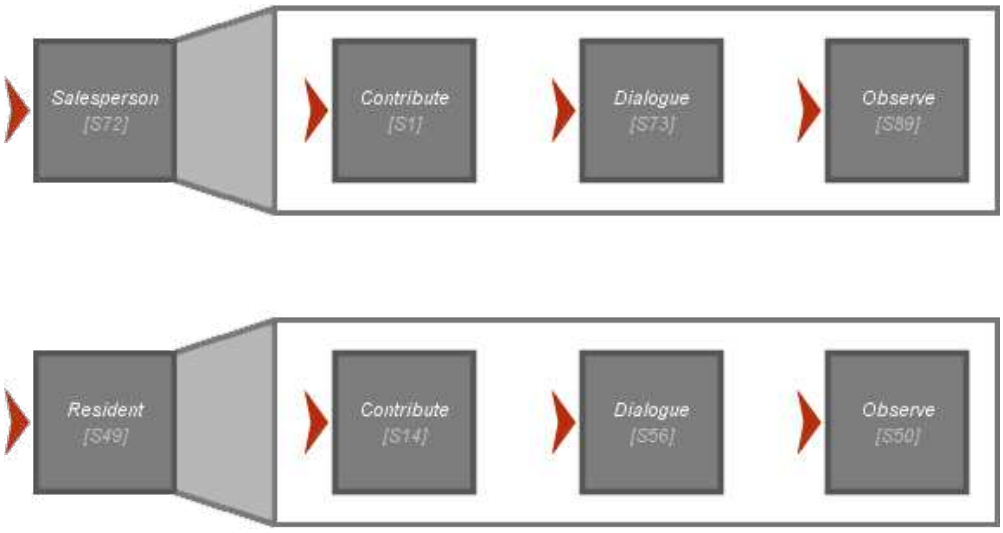


Figure 9.4: The major processes controlling each agent's behavior, hierarchically encapsulated in the *Interaction* supernode.

Minimum Necessary Information

One crucial prerequisite for interleaving dialogue contributions is to know when a participant has heard or said enough to proceed. Therefore, the dialogue manager must be able to detect and reason about the *minimum necessary information (MNI)* [26], as was explained in sections 5.2.1 and 7.2.1.

In this prototype, the dialogue was entirely pre-scripted, so the straightforward solution was to insert bookmarks into the spoken text and associate them with a unique content identifier. Visual SceneMaker supports this via markers that are inserted into the scene script. At runtime, the *InterruptibleExecutor* extracts these markers, stores them separately, and replaces them with unique bookmark tags in the syntax required for the respective agent's text-to-speech engine. When an agent reports a bookmark event, the marker associated with its identifier is retrieved and the action described by this marker is executed.

For registering the MNI, a custom action named "minInfo" was defined. Any given parameters were interpreted as semantic content identifiers. So whenever a "minInfo" marker is triggered, these identifiers are reported to the *SharedInfoExexcutor*. This executor provides a method *waitForInfo* that returns *true* if the requested information is already present or stalls until this information arrives.

Figure 9.5 shows how this function is combined with the *interruption edges* offered by Visual SceneMaker. As soon as a supernode is reached, Visual SceneMaker starts evaluating the function on the interruption edge that connects it to its successor. So, once the requested information is registered with the Shared Information Board, whatever happens in that supernode is interrupted

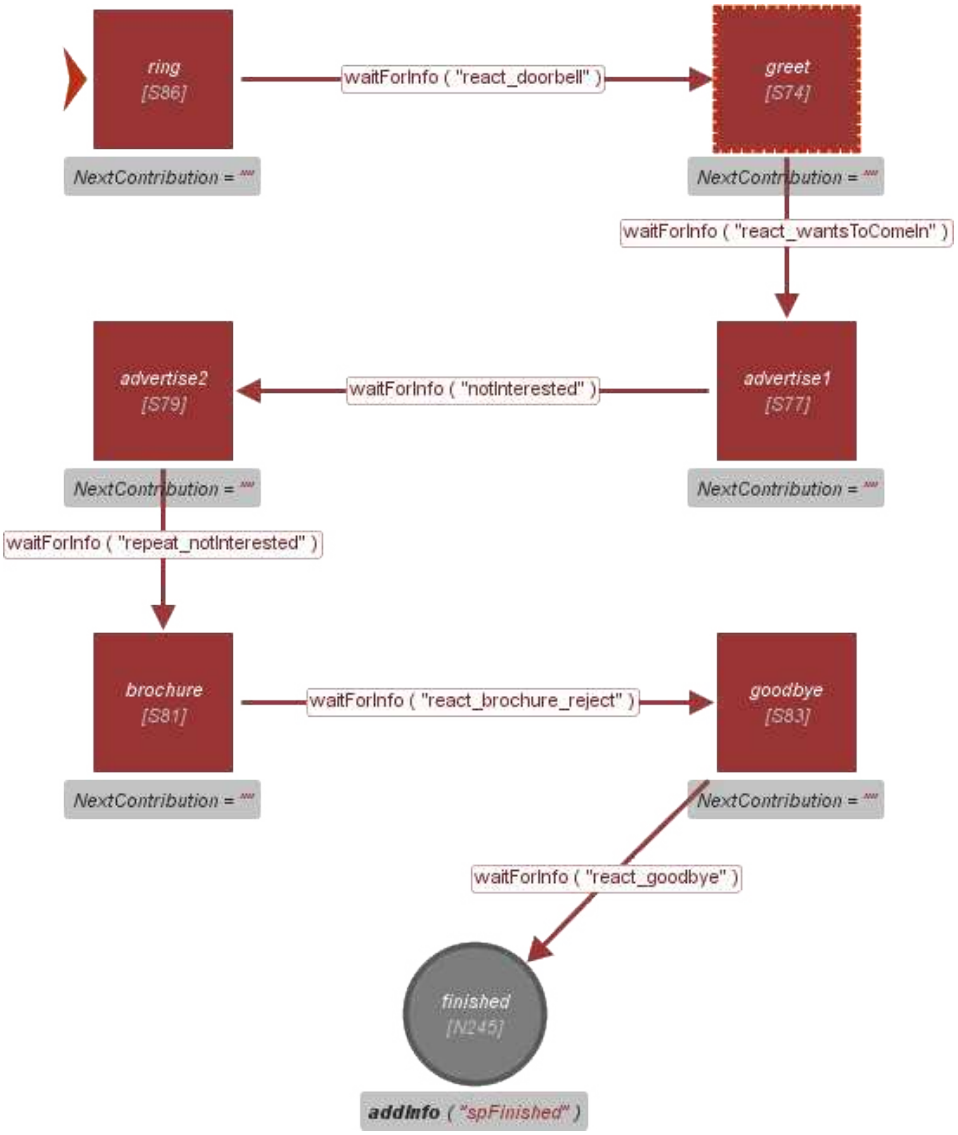


Figure 9.5: Dialogue flow for the *Salesperson* character.

and the state machine transitions to the character's next turn.

Figure 9.6 shows one of those turn supernodes. It holds three phrases that the agent will try to speak, each of which is again represented by its own state machine. At the end of the chain is a waiting node that, if necessary, pauses the dialogue flow until the proper response from the interlocutor becomes available. When a phrase node is reached, the variables *NextContribution* and *AwaitedMNI* are set so that the *Contribute* state machine will know which utterance will come next, and how to tell when enough of said utterance has been communicated.

The *Contribute* state machine then uses the variables set by its sibling processes to dispatch or interrupt the current speech command at the appropriate

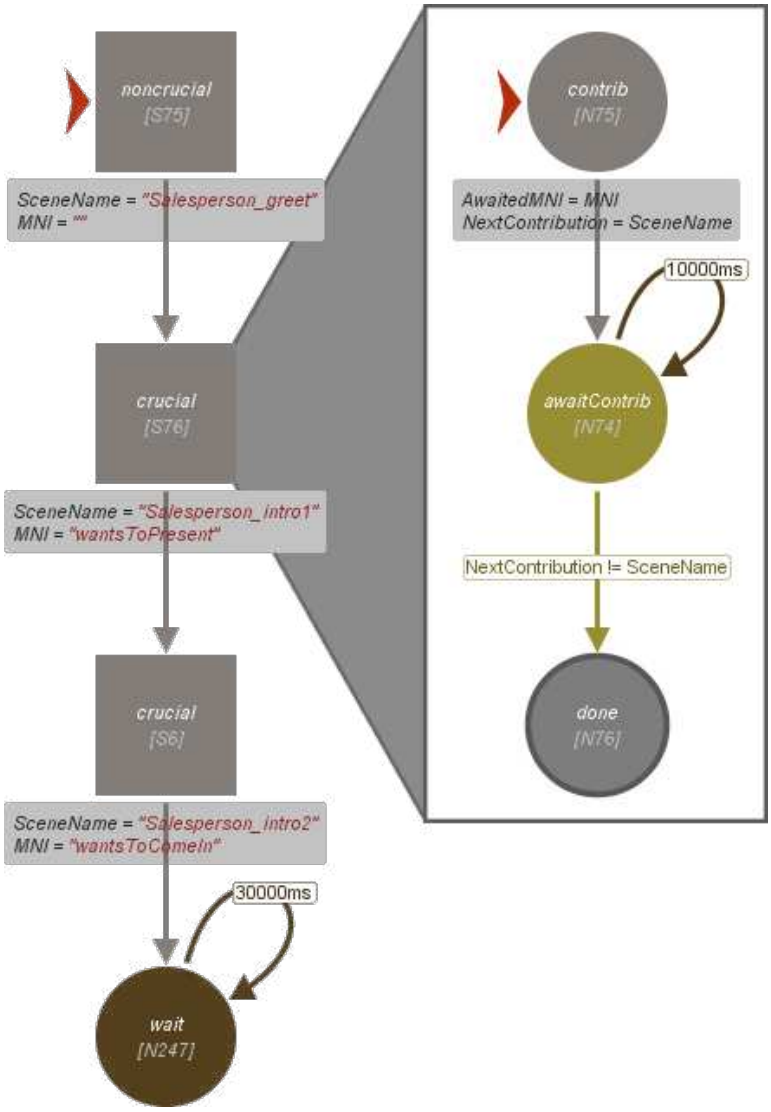


Figure 9.6: Example for an agent's turn. Every phrase is modeled as a supernode that sets the name of the associated scene and, if applicable, the identifier of its semantic content. The box on the right shows the contents of a phrase supernode.

time. When the *Dialogue* machine proposes a new contribution by setting those variables, the *Contribute* machine waits for the influence diagram's permission to take the floor, sends the respective speech command to the agent, and if necessary, interrupts the speech command if the decision changes before it could finish. After that, the Shared Information Board is checked to see whether the message has been transmitted successfully. If that is not the case, the state machine tries to re-issue the speech command at the next opportunity. Otherwise, it clears the *NextContribution* variable and waits for the next contribution to become available.

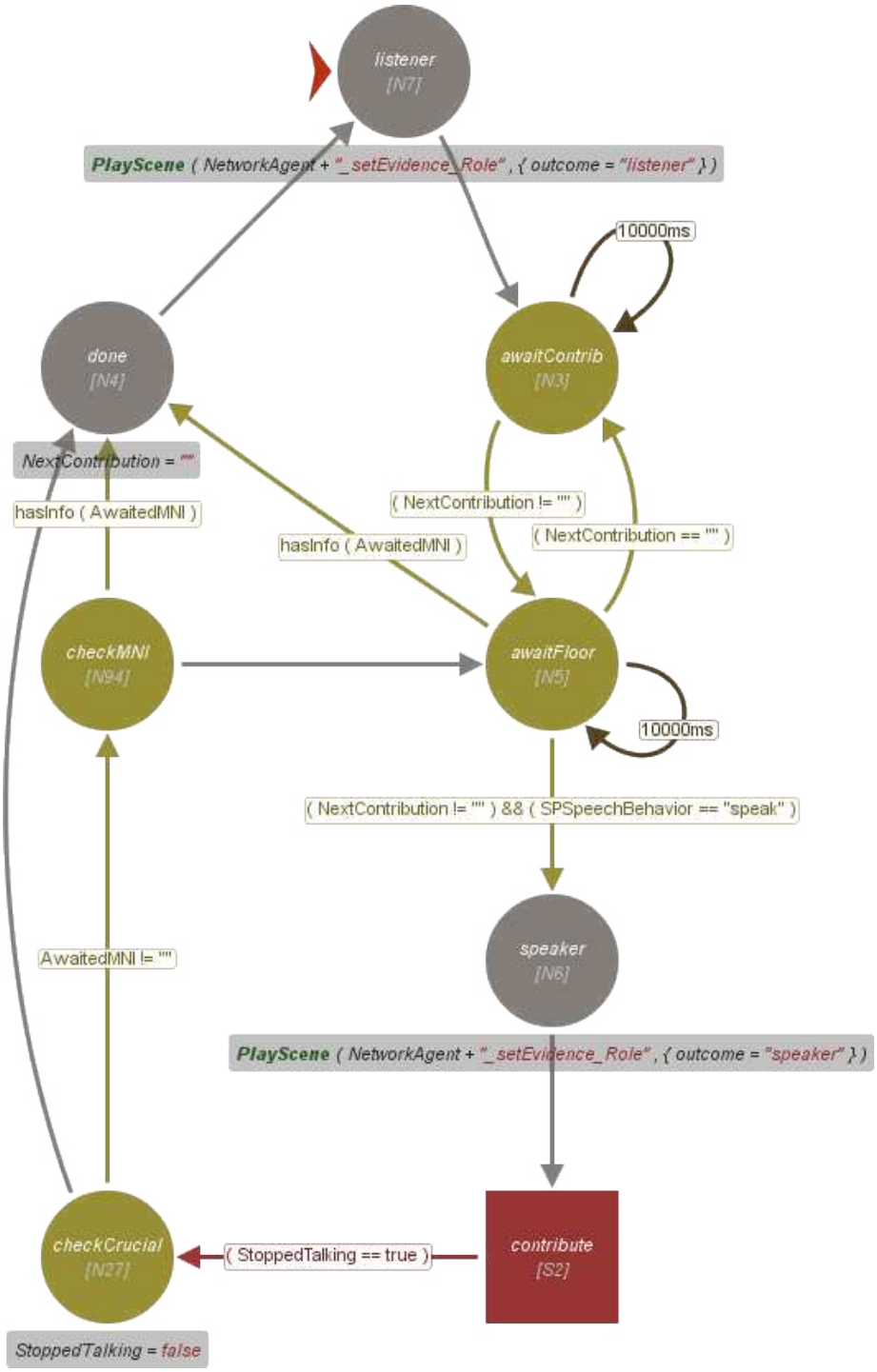


Figure 9.7: The *Contribute* state machine that handles the starting, stopping and repeating of speech commands.

```
[
   type: event,
   time: 39749274581,
   dist: 0,
   life: 0,
   conf: 1.0,
   name: 'Salesperson',
   mode: voice,
   data: 'true'
]
```

```
[
   type: event,
   time: 39749276352,
   dist: 0,
   life: 0,
   conf: 1.0,
   name: 'Salesperson',
   mode: voice,
   data: 'false'
]
```

Figure 9.8: Typed feature structures representing the voice activity events. *Left:* Event raised at the beginning of the speech output. *Right:* Event raised after the end of the speech output.

When the *NextContribution* variable is reset, the *Dialogue* process knows that the current phrase has been spoken and proceeds to the next phrase node. Because not all phrases need to be repeated, it is possible to leave the MNI variable blank and thus make the *Dialogue* process skip non-crucial phrases after the first failed attempt. In contrast, phrases that advance the dialogue must be associated with a unique MNI identifier that matches both the one expected by the conversation partner and the one found in the scene script.

Agent Perception

The agents' voice activity is treated like sensor data from a human user to facilitate the later addition of input processing. Therefore, each time an agent starts or stops speaking, a sensor event is generated and stored in the Prolog fact base as done in the work by Mehlmann et al. [83, 84]. Two examples of typed feature structures representing such events are shown in figure 9.8.

Two dedicated subsections of the state machine, found in each agent's *Observe* supernode, are responsible for extracting these events from the fact base, processing them, and updating the respective agent's Bayesian network accordingly. These subsections consist of three parallel processes, shown in figure 9.9.

The most basic process, *Detect Voice*, contains only one node with a timed edge looping back to it after 250 ms. Every time this node is reached, it queries the fact base for an event with the *voice* modality and the other participant's name. If such an event exists, it is removed from the fact base, and the event data - a boolean value indicating the presence or absence of speech - is stored in a dedicated variable.

The second process, *SpeechState*, repeatedly checks for changes in this variable's value. On detecting such a change, the boolean value is translated to the

Figure 9.9: Subprocesses modeled in the "Observe" state machine section.

chance node outcomes *silent* respectively *speaking*, and the observed evidence is set in the agent's own Bayesian network.

Finally, the process *SpeechStDur* tracks the duration of the other participant's current speech state. When that variable changes, this process moves through a chain of nodes that set the duration evidence in the Bayesian network to *very short*, *short*, *long* and *very long*, respectively. The nodes are connected via timed edges so that the duration of each interval can be configured easily.

Influence Diagram

The influence diagram was realized using the SMILE library and the GeNIe editor[6]. The *BayesianNetworkExecutor* implemented for this prototype loads the network file created in GeNIe and provides an interface for updating the evidence at its chance nodes. Furthermore, every time the network is updated, the outcome of the decision node for the speech behavior is written to a variable within the SceneMaker project so that the agent's behavior can be adapted accordingly.

Figure 9.7 shows how said variable - in this case, "SPSpeechBehavior" for the "Salesperson" character - is used as an additional condition for delaying the speech command within the "Contribute" section of the agent's state machine. Figure 9.10 shows the subsection which handles the actual speech command, encapsulated in another node named "contribute". When this node is reached, two parallel processes start within it. One plays the scene whose identifier was stored in the "NextContribution" variable and sets the "StoppedTalking" flag to true when the command execution has finished. Meanwhile, the other process waits for the network's decision to change to *wait*, at which point a *stopSpeech* command is sent to the agent. When the agent stops speaking, the original speech command registers as finished, triggering the transition to the node that sets the "StoppedTalking" flag. As soon as this flag becomes "true",

[6]both by BayesFusion, LLC, and available free of charge for academic teaching and research use at http://www.bayesfusion.com/

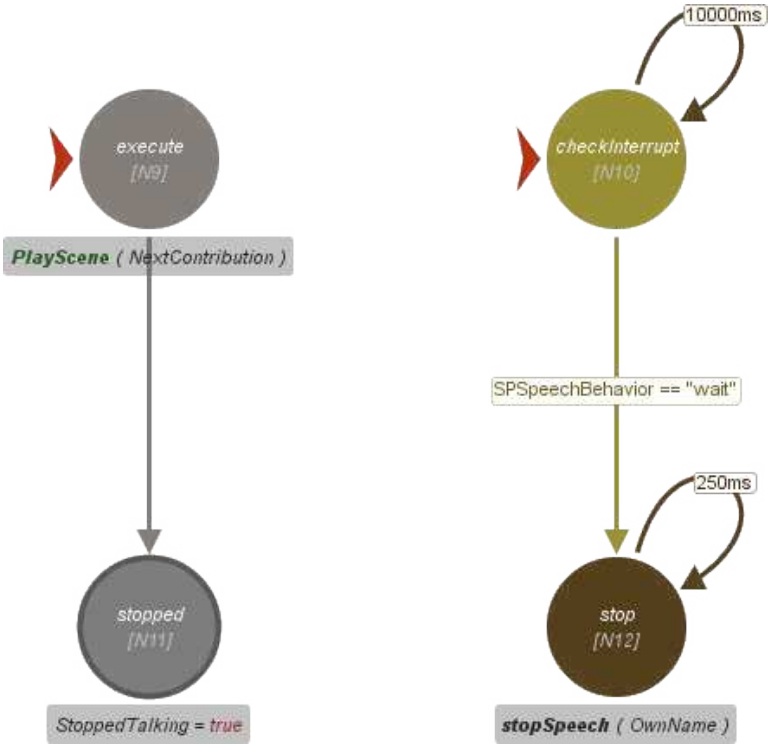


Figure 9.10: Subsection of the "Contribute" state machine section which executes the speech command in an interrubtible manner.

the surrounding node is exited via an interruption edge (figure 9.7). This way, the cycle continues regardless of whether the speech command finished normally or prematurely.

A simple visualization was implemented to debug the activity within the influence diagram. In addition to each node's type and current values, the color and screen coordinates - as set in the GeNIe editor - were easily accessible through the SMILE API. So were the parent-child relationships. Thus, it was straightforward to replicate the network's appearance in a dedicated window. Furthermore, this interface allows for setting the outcome of any chance node manually at runtime.

Speech Obfuscation

Finally, a function was implemented to render the spoken text unintelligible on demand. This was done to avoid biasing observers through the semantic context, as will be explained in section 9.4 regarding the prototype's evaluation.

This obfuscation was eventually implemented as a function in the *InterruptibleExecutor* so that it could be reused for any agent type connected to the Visual SceneMaker, not just the PseudoBots.

The consonants and vowels in the text are systematically re-ordered so that the duration of each word remains the same, but its sound no longer makes sense. Numeric and special characters are ignored. At the time the text is scrambled, the *minInfo* markers in the scene script have already been converted to the generic format of a dollar sign followed by a number, and therefore the MNI bookmarks remain in the same place. This approach ensures that the semantic information appears on the Shared Information Board at the same time it would appear without the scrambling.

9.4 Perception Study

A perception study was conducted online to evaluate this model and ensure that it can generate the desired impression of the agent's personality and interpersonal attitude. This section will describe the experimental procedure and discuss its results.

9.4.1 Hypotheses

The non-interactive prototype was intended as a first proof of concept, and therefore, the primary goal was to replicate the results found in related works. Objectively, the decisions of the influence diagram lead to shorter pauses and longer overlaps when an agent was configured as extraverted or dominant. Now, it was necessary to test whether human observers would interpret these behavior patterns in a similar manner to those examined by ter Maat et al. [130] and Glas et al. [47]. Moreover, the study was meant to confirm the theoretical relationship between the Big Five factors and the Interpersonal Circumplex as explained in section 3.2. Therefore, the hypotheses for this prototype were as follows:

- Hypothesis 1: An agent's Extraversion score will be higher when it is configured as *extraverted*.
- Hypothesis 2: An agent's Status score will be higher when it is configured as *extraverted*.
- Hypothesis 3: An agent's Agreeableness score will be lower when it is configured as *extraverted*.

Hypothesis 1 was meant to verify that the Extraversion parameter was correctly reflected in the character's speech timing. The other two hypotheses were based on the relationship between the personality and interpersonal attitude dimensions (see section 3.2.3). Therefore, higher Extraversion was

expected to imply higher Status as well. Since the timing in this prototype is only based on the Status and the same Status level can result from different personality configurations, Agreeableness was expected to be affected.

9.4.2 Experimental Validation

This prototype did not feature any user input, so it was decided to conduct an online survey based on video recordings. This made it possible to obtain a large number of ratings within a short time frame.

Stimuli Preparation

The Extraversion levels of both agents were chosen as the independent variables, each with two levels *introverted* and *extraverted*. Therefore, 2×2 videos were recorded to represent all possible combinations.

Two PseudoBot agents were set up in neighboring windows, oriented to face each other. To avoid biases by their appearance, they used the same 3D model and displayed the same neutral gray color. Additionally, both agents' TTS engines used the same voice with identical pitch range and volume, so that the ratings would not be influenced by apparent differences in gender, age, or ethnicity. The only difference between them was the audio channel on which the respective agent could be heard. To help subjects distinguish which agent was speaking at any given time, the left agent's voice was sent to the left channel while that of its right counterpart was sent to the right side.

As for the spoken content, a dialogue script was prepared to generate naturally timed phrase lengths and interruptions. The conversation was between a door-to-door salesperson and a resident who was not interested in their product. This scenario was chosen because it was short but still provided multiple opportunities for both parties to try interrupting each other. Table 9.8 shows the script for both agents, with each MNI that triggers the next turn marked by an asterisk.

However, using intelligible speech would have carried the risk of biasing the study subjects through the scenario itself. For example, some people might believe the resident to have a more legitimate reason to control the conversation than the salesperson intruding on their home. Alternatively, some might think that salespersons are "naturally" pushy or good at persuading people because this makes them successful in their job. Therefore, the semantic context was obfuscated through the method described in section 9.3.2. This resulted in the agents talking in gibberish while using the same timing as in the original dialogue.

	Salesperson	**Resident**
1a	Ring ring*!	
2a		Hello*?
3a	Good day!	
3b	I am from the company Dirt-B-Gone and I would like to present our newest vacuum cleaner to you.	
3c	May I come in* for a minute?	
4a		Uh,
4b		I don't know*.
4c		Actually I am quite satisfied with my old vacuum cleaner.
5a	Believe me,	
5b	Compared to our Slurp 380* your old vacuum cleaner will look like a stone age relic.	
6a		No thank you*.
6b		I am not interested.
7a	The Slurp 380* is the world's first vacuum cleaner using the revolutionary Piranhanado technology!	
7b	No matter how fiercely the dirt digs its teeth into the carpet,	
7c	our Slurp 380 is stronger.	
8a		I told you I am not interested*!
8b		I don't need your Slurp.
8c		Please leave.
9a	May I at least leave this brochure* with you?	
10a		No*.
11a	Alright.	
11b	Nevermind then.	
11c	Have a nice* day!	
12a		Whatever.
12b		Bye bye.

Table 9.8: The dialogue script used for generating the video stimuli. The * marks the end of the MNI which will cause the interlocutor to move on to the next turn.

The final video stimuli were trimmed to start with phrase 3c (see table 9.8), removing the strictly sequential exchange at the beginning. The final clip durations were 0:33 minutes for *(extraverted-extraverted)*, 1:03 minutes for *(extraverted-intoverted)*, 0:50 minutes for *(introverted-extraverted)* and 1:15 minutes for *(introverted-introverted)*.

In addition to the stimuli, a short sound test video was recorded. In this clip, the agents were speaking normally to inform the viewer which of them should be heard on the left respectively right side. After that, they asked the viewer to adjust the volume accordingly and recommended the use of headphones.

Questionnaire Design

The questionnaire was implemented using LimeSurvey[7]. It was made available in both German and English.

After some general information about the study, the questionnaire started with the sound test video and asked the participant to write down the fourth sentence that was spoken to verify that they could view the videos as intended.[8]

For demographic context, the participants were asked about their age group, gender, first language, and occupation. Their previous experience with computer-controlled characters was assessed by asking how much contact they had had with three different agent types, namely video game characters, voice assistants, and social robots. Possible answers were "no experience at all", "have seen it in action", "have used it myself", and "use it regularly myself".

The questionnaire items for rating the agents' personalities were selected from the BFI-10 by Rammstedt and John [104], available in both English and German. For dominance, however, validated items were harder to find. For instance, the IPIP-IPC by Markey and Markey [79] measures the octants rather than assessing the status dimension directly. Furthermore, many statements such as "speak softly" or "speak loudly" did not apply to these stimuli, while items such as "let others finish what they are saying" or "do most of the talking" were too focused on the surface behavior. Since the study's goal was to confirm the observers' beliefs about characters displaying said behavior, this would have led to circular reasoning.

Other questionnaires for measuring dominance were not applicable, either. Most of those were created for assessing interpersonal dynamics in romantic relationships, for example, to detect abusive tendencies in one of the partners. Therefore, those questionnaires consisted of items like physical or emotional

[7] https://www.limesurvey.org

[8] This initial test was added after the first few subjects took part, following a suggestion by Birgit Lugrin.

abuse, reflecting circumstances that were completely unrelated to the conversation topic.

Finally, as an incentive for completing the survey, participants were offered the chance to enter a lottery for one of three Amazon gift cards worth 10 € each. If they chose to do so, they could provide their email address at the end of the questionnaire.

The survey items in German and English can be seen in appendices A.2.2 and A.2.3, respectively.

Recruitment of Participants

10 Din-A4 posters were put up in various locations on the campus of the University of Augsburg, such as outside the cafeteria or on the doors leading to the chair of Human-Centered Multimedia. 62 flyers advertising the survey were placed in 5 different locations, mostly in the computer science building but also in the mathematics building. 70 more flyers were handed out to people all over the campus. A similar advertisement was made in a journal entry on the artist platform DeviantArt. The design of the posters, flyers, and the DeviantArt journal can be seen in appendix A.2.1.

Present colleagues at the chair were invited to the survey in person, while former colleagues and current cooperation partners were contacted via email. For instance, an invitation to the survey was sent out via the mailing list of the ForGenderCare project. Furthermore, a substantial number of survey answers could be obtained using the participant recruitment system of the University of Würzburg[9], where students received course credit for taking part.

9.4.3 Results

The survey was completed by 116 participants (44 male, 70 female, 2 no answer). The majority (73.3%) was in the age group from 20 to 29, and almost all of them (94%) named German as their first language. Most of the participants (79%) were university students, mainly from subject areas related to computer science or media communications. Therefore, the familiarity with computer-controlled agents was relatively high. Most had already interacted with video game NPCs and voice assistants and at least seen social robots in action.

The questionnaire items concerning each measured trait - Extraversion, Agreeableness, and Status - were combined into a single score for the respective trait for each agent and condition. 2x2 repeated measures MANOVA were performed to determine whether each agent's configured Extraversion influenced

[9]Many thanks to Birgit Lugrin for providing this opportunity!

Configured Extraversion		**Perceived Extraversion**			
Left Agent	Right Agent	Left Agent		Right Agent	
trueEL	*trueER*	Mean	SD	Mean	SD
introverted	introverted	3.71	0.65	2.53	0.79
introverted	extraverted	3.28	0.62	3.62	0.69
extraverted	introverted	3.92	0.64	2.28	0.72
extraverted	extraverted	4.02	0.51	3.58	0.73
Configured Extraversion		**Perceived Status**			
Left Agent	Right Agent	Left Agent		Right Agent	
trueEL	*trueER*	Mean	SD	Mean	SD
introverted	introverted	3.68	0.69	2.55	0.70
introverted	extraverted	2.61	0.86	3.59	0.83
extraverted	introverted	3.84	0.68	2.46	0.76
extraverted	extraverted	3.44	0.67	2.84	0.83
Configured Extraversion		**Perceived Agreeableness**			
Left Agent	Right Agent	Left Agent		Right Agent	
trueEL	*trueER*	Mean	SD	Mean	SD
introverted	introverted	3.46	0.61	3.48	0.62
introverted	extraverted	3.58	0.69	2.05	0.61
extraverted	introverted	3.39	0.61	3.40	0.59
extraverted	extraverted	2.23	0.54	2.37	0.75

Table 9.9: Results of the perception study. Perceived traits range from 1.0 (very low) to 5.0 (very high).

its perceived personality and interpersonal attitude. Pairwise comparisons were based on the estimated marginal means with Bonferroni correction.

In the following, *trueEL* will denote the left agent's configured Extraversion while *trueER* will denote that of the right agent.

Perceived Extraversion

There was a significant main effect of *trueEL* on the left agent's perceived Extraversion ($F(1.0, 115.0)=112.97$, $p=0.000$). When set to *extraverted*, it received a higher score ($M=3.97$, $SD=0.58$) than when it was *introverted* ($M=3.50$, $SD=0.67$, $p=0.000$). For the right agent, there was a significant main effect of *trueER* on its perceived Extraversion ($F(1.0, 115.0)=223.13$, $p=0.000$).

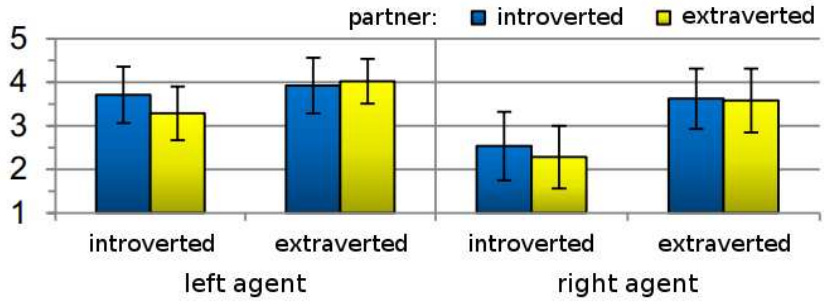


Figure 9.11: Extraversion scores for the two agents, ranging from 1 (very introverted) to 5 (very extraverted).

When set to *extraverted*, it received a higher score (M=3.60, SD=0.06) than when it was *introverted* (M=2.41, SD=0.76, p=0.000). Therefore, Hypothesis 1 was confirmed.

There also was a significant main effect of the left agent's configuration on the score of the right agent (F(1.0, 115.0)=8.17, p=0.005) and vice versa (F(1.0, 115.0)=10.51, p=0.002). In both cases, the difference between the Extraversion scores was more pronounced when the conversational partner was configured as *extraverted*. The effect size was small for the left agent when *trueER* was *introverted* (Cohen's d = ±0.33), but large when *trueER* was *extraverted* (Cohen's d = ±1.29). For the other agent, the effect size was large in both cases (Cohen's d = ±1.46 versus ±1.79)

One possible explanation is that an *extraverted* partner is required to see an agent's reaction to being interrupted, which then makes it easier to spot the difference in behavior. Overall, the Extraversion score was higher for the left agent, which can be explained by the fact that it always initiated the conversation and, if not interrupted, had more text to say.

Perceived Status

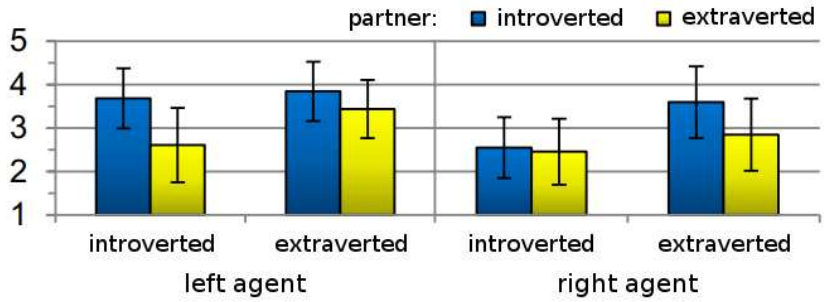


Figure 9.12: Status scores for the two agents, ranging from 1 (very submissive) to 5 (very dominant).

There was a significant main effect of *trueEL* on the left agent's perceived Status (F(1.0, 115.0)=76.01, p=0.000). When set to *extraverted*, it received a higher score (M=3.64, SD=0.71) than when it was *introverted* (M=3.14, SD=0.95, p=0.000). For the right agent, there was a significant main effect of *trueER* on its perceived Status (F(1.0, 115.0)=74.73, p=0.000). When set to *extraverted*, it received a higher score (M=3.22, SD=0.91) than when it was *introverted* (M=2.50, SD=0.73, p=0.000). Therefore, Hypothesis 2 was confirmed.

As with the Extraversion score, there was a significant main effect of *trueER* on the Status score of the left agent (F(1.0, 115.0)=115.26, p=0.000), and the difference in the score was more pronounced when the other party was *extraverted* (Cohen's d = ±1.08 versus ±0.24). There was also a main effect of *trueEL* on the right agent's Status score (F(1.0, 115.0)=53.01, p=0.000). However, in that case, the effect was stronger when the left agent was *introverted* (Cohen's d = ±1.37 versus ±0.49). This may be because the left agent was perceived as having a higher Status in general, which is in line with its higher Extraversion score and the dependency between those two dimensions (see section 3.2.3). This, in turn, could mean that it overshadowed the right agent's behavior differences when it was set to *extraverted*.

Perceived Agreeableness

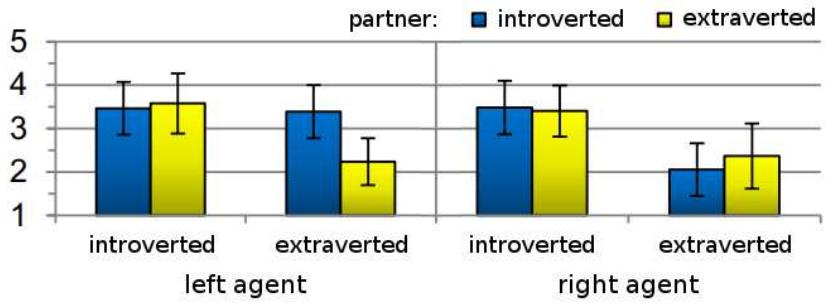


Figure 9.13: Agreeableness scores for the two agents, ranging from 1 (very disagreeable) to 5 (very agreeable).

There was a significant main effect of *trueEL* on the left agent's perceived Agreeableness (F(1.0, 115.0)=182.22, p=0.000). When set to *extraverted*, it received a lower score (M=2.81, SD=0.82) than when it was *introverted* (M=3.52, SD=0.65, p=0.000). For the right agent, there was a significant main effect of *trueER* on its perceived Agreeableness (F(1.0, 115.0)=341.33, p=0.000). When set to *extraverted*, it received a lower score (M=2.21, SD=0.70) than when it was *introverted* (M=3.44, SD=0.60, p=0.000). Therefore, Hypothesis 3 was confirmed.

Again, there were significant main effects of *trueER* on the left agent's Agreeableness score (F(1.0, 115.0)=102.81, p=0.000) and of *trueEL* on the right agent's Agreeableness score (F(1.0, 115.0)=5.36, p=0.022). For the left agent, the effect was only notable when the other party was *extraverted* (Cohen's d = ±2.16 versus ±0.12), whereas for the right agent, the effect was stronger when the left agent was *introverted* (Cohen's d = ±2.34 versus ±1.54). This matches the results for the Status score and confirms that higher Status implies lower Agreeableness and vice versa (see section 3.2.3).

Axis Rotation

One more thing to confirm was the relationship between Extraversion, Agreeableness, and Status. Therefore, a search was performed for the angle that would best explain the status score over the 928 agent ratings obtained in the study. For every agent rating in the questionnaire answers, the status value was predicted by mapping the agent ratings to the range [-1.0;+1.0] and then rotating the *(Agreeableness, Extraversion)* vector. Afterward, values for α in the range of $[-45°; -20°]$ were tested to see which angle would minimize the mean square error of the predicted Status value compared to the measured value.

$$\begin{aligned} PredictedStatus_i(\alpha) &= sin(\alpha) * (Agreeableness_i * 2 - 5) * 0.2 \\ &\quad + cos(\alpha) * (Extraversion_i * 2 - 5) * 0.2 \\ MeasuredStatus_i &= (Status_i * 2 - 5) * 0.2 \end{aligned}$$

$$\underset{\alpha \in [-45°; -20°]}{\arg\min} \frac{\sum_{i=1}^{928} (PredictedStatus_i(\alpha) - MeasuredStatus_i)^2}{928}$$

The mean and standard deviation of the squared error are plotted in figure 9.14, and table 9.10 lists the minima that were found within this range. The angle used for calculating the conditional probabilities was -37.5°, and indeed, there is a minimum only 0.5° away from it. However, the fact that the minima in the mean and standard deviation repeat periodically indicates that the angle cannot be analyzed properly without knowing the Affiliation value that belongs to the respective agent rating.

Additional Comments

10 participants stated additional observations via the comment fields. The detailed comments and their translation to English can be found in appendix A.2.4.

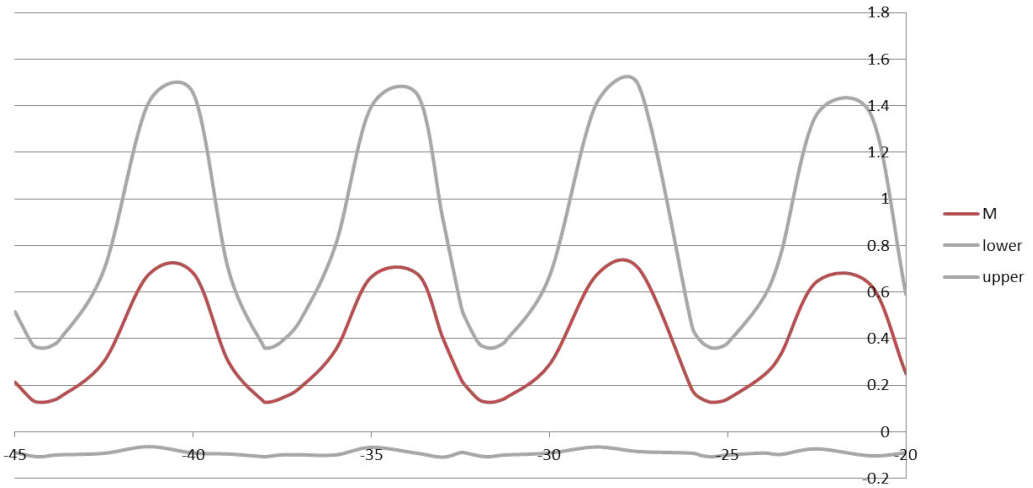


Figure 9.14: The mean squared error (center line) and associated standard deviation (upper and lower contours) depending on the chosen α.

α	M	SD
-25.39812	0.12615	0.233123
-31.68130	0.12615	0.233123
-37.96449	0.12615	0.233123
-44.24767	0.12615	0.233123

Table 9.10: Angles that minimize the mean squared error in predicting the Status rating from the Extraversion and Agreeableness ratings.

Some described how long the agents paused before speaking and how often they interrupted the other one. They all observed correctly. One person also noticed that the left agent initiated the conversation while the right one merely reacted. As stated above, this may explain the left agent's higher Extraversion and Status.

Others referred to the general tone of voice. One participant found that the agents sounded rather negative, as if they were arguing. Regarding the *extraverted* × *extraverted* conversation, another concluded from the monotonous speech that it was not the type of passionate discussion in which overlapping speech was tolerated and even desirable. A different participant wondered whether those two agents were talking to each other or conducting unrelated phone calls, hinting at a lack of interest implied by the overlaps.

Two comments pointed out that aspects like the degree of control were hard to determine from the meaningless sounds. One participant gave examples of situations in which overlaps were common and even desirable, which confirms

that a believable turn-taking model needs to consider many context factors (see sections 3.4 and 5.2).

Finally, one person recognized the synthetic voice as that of an American male and drew attention to the strong gender bias induced by this. Said bias was, in fact, the reason why both agents had been given identical voices so that they would not be rated differently from each other based on their apparent gender or culture. However, future studies should certainly explore the effects of different voices.

9.4.4 Discussion

Overall, the study confirmed the findings from related works on which this prototype had been built. The timing modifications generated through this computational model affected the agent's apparent personality and attitude in a way that was comparable to the results of ter Maat et al. [130] or Glas et al. [47]. The study results also add further support for the theoretical connection between Extraversion, Agreeableness, and Status. However, several limitations became apparent.

Semantics and Additional Modalities

This prototype covers only the raw speech activity and thus barely scratches the surface of all those clues humans employ for communicating their intentions. It considers neither semantic information, such as the disruptiveness of a question [47], nor the gaze signals that have been found to play a major role (see section 3.4.2).

Study participants commented on their difficulties in judging an agent's personality or attitude from the timing alone. Consequently, future experiments should use clearly intelligible speech rather than the scrambled sounds used in this study.

They also pointed out other factors, such as the monotonous or hostile-sounding synthetic voice or the fact that one agent initiated the conversation. These factors should be controlled for in future experiments.

Simplified Goal Definition

While it was useful for the first proof of concept, the way the interaction goal had been defined was not ideal. Although the utilities and their connection to the world state did indeed produce behavior consistent with existing research, the number of influence factors on which the goal depended still made it difficult to understand their interplay.

Later attempts to specify additional goals revealed that *Exert Control* was highly abstract, as it related an arbitrarily defined, subjective scale ("how much control does this character appear to have") to heuristically adjusted "desirability" values for the given timings. This "desirability", while in line with the rise, fall, and turning points of the costs and benefits, was not actually based on a well-defined cost function.

According to Abbas' book "Foundations of Multiattribute Utility" [1, p. 79], using arbitrary scales is a common mistake. Additionally, people frequently treat the *probability* of obtaining a given prospect as an attribute of the prospect itself [1, p. 156]. These widespread yet incorrect practices lead to models that appear to make appropriate decisions on the surface, but when they fail, it may be difficult to understand what went wrong. This is exactly the kind of problem that the approach described in this thesis is supposed to avoid.

Consequently, the core model needs to be revised before adding the new goals. Later influence diagrams have to represent the influence factors and their effects in a more detailed and rigorously structured way.

Scalability

Conditional independencies between subsections of the model help to reduce its complexity and, therefore, the necessary amount of computation. Nevertheless, the influence diagram needs to be accessed by multiple parallel processes for the different modalities, such as speech and gaze. Both operate on a timescale of split-seconds, so computational efficiency and proper thread synchronization are crucial.

Later tests showed that the efficiency of the surrounding dialogue setup can contribute to the delays, making it hard to see whether the behavior model works as intended. For example, on the older laptop[10] used for this prototype, the MaryTTS engine could take up to two seconds to generate the audio output for a longer sentence. The way the state machines are designed can also interfere with the behavior timing. Interaction designers need to be careful not to override the influence diagram's decisions by, for example, forcing the agent to stop speaking as soon as a topic shift is detected.

Interactive Human-Agent Dialogue

Many open challenges fall into the research area of incremental input and output processing. For instance, the end of the MNI needs to be detected at runtime. For small domains, such as the salesperson dialogue in this thesis or

[10]The specifications of the laptop are listed near the end of this thesis, in table 12.1.

the "Simon says" game in Chao's work [26], relevant keywords can be marked by hand. Complex domains such as negotiation training simulations [38, 137] require learning the MNI from large corpora of user utterances and synonymous phrases. One common approach is to look for known concepts and entities in the already spoken text and let the interaction advance once the necessary slots are filled [38, 137].

As for output generation, it is possible that the agent is not yet ready to respond when the turn should be taken. In our prototype, the influence diagram's decision does not force the agent to speak at that point but rather adds a delay if the opposite is true. A suitable extension would be to take the turn and employ turn-hold signals such as filler words [17, 122] and gaze aversion [4, 122] while waiting for the content generation to finish.

9.5 Conclusion

This chapter described a non-interactive proof-of-concept prototype for controlling turn-taking timing with an influence diagram. It showed how such an influence diagram can be constructed and how it can be integrated with a dialogue application.

In preparation for the interactive setup, the agents talking to each other were not allowed to access their partner's intentions directly. Just like humans who cannot read each other's minds, the agents had to rely on the perceivable voice activity to infer the other participant's current role through the mechanisms of the Bayesian network. This inference then allowed them to choose the action that was most likely to fulfill their communicative goal.

For this simplified prototype, the focus was on the agent's desire to control the interaction, which depends on its personality and, consequently, its attitude toward the other participant. This goal was chosen because its connection to extraverted and dominant personalities was most salient in psychological research and related works from computer science.

The prototype was then validated in an online study using video clips of short conversations generated with the developed application. The results confirmed that the agents' behavior leads to the desired Extraversion perception. Its effects on the perceived Status and Agreeableness are in line with existing psychological theories. However, the survey also confirmed that turn-taking depends on many more factors than those implemented so far, which further stresses the need for extensible, adaptable behavior models.

While creating the influence diagram, it also became apparent that a more transparent approach is needed for defining an action's utilities with regard to a specific goal. This issue will be addressed in the next chapter.

Chapter 10

Human-Agent Conversation

10.1 Introduction

Simulating conversation between agents is useful as a proof of concept, but the actual goal is to have these agents interact with humans. For this, it becomes necessary to deal with the semantics of the exchanged messages.

The non-interactive prototype confirmed that varying an agent's speech timing lead to the expected differences in personality perceptions. However, it also showed that study participants felt unsure about their judgment when the conversation topic was hidden behind gibberish. Together with the need for detecting the MNI, this called for a more complex setup with actual speech recognition and intent parsing.

In addition, one crucial modality has been ignored so far. In human communication, turn-taking behavior is tightly coupled with the gaze behavior of the interaction partners. While it is possible to coordinate speaking activity without seeing each other, humans use this information to understand the other party's intentions more accurately. The related patterns have been shown to apply to human-agent interaction as well. (See sections 3.4.2 and 5.3.1 for more information.)

The non-interactive prototype had already been set up so that replacing one agent with a human was relatively straightforward. However, it was still necessary to find suitable input recognizers and extend the behavior model to support a more realistic conversation. Evaluating the interactive setup also turned out to be challenging, and while that task was not entirely completed, several valuable lessons could be learned on that topic.

This chapter describes the development of an interactive, multimodal dialogue setup. First, the modeling approach of the non-interactive prototype is

revised, and the model is extended to cover more personality traits and more systematically defined goals. The next section describes the implementation, especially the changes made to improve the support for different modalities and the semantic information exchanged through them. There will also be subsections on procedural gaze animation and the recognition of user input. The section after that will describe an initial approach for evaluating that setup before another section will discuss the insights that were gained in the process. The chapter ends with a summary.

10.2 Influence Diagram

The main challenge for the revised influence diagram was to identify concrete prospects that could be compared to the agent's goals. The solution found in this thesis was to reduce the abstract, personality-based motivations to the low-level goals of sending and receiving information. The personality traits and interpersonal attitude then shape the agent's behavior by determining how the agent prioritizes those.

Furthermore, a distinction was made between a goal that the agent tends to pursue in general and one that it would pursue in a given dialogue context. For example, receiving information has a high base priority for a curious agent that cares about its interaction partner, but can be overshadowed when an urgent message gets delayed for too long, or become irrelevant when the interlocutor is not saying anything.

The earlier model's decision between acting and waiting was replaced with a more general choice between paying attention to the partner or oneself. This attention model allowed for a transparent representation of the agent's cognitive processes, especially the fulfillment of its information-related needs. This approach was used for both considered modalities, speech and gaze.

This section takes a closer look at the influence diagram that was developed for the interactive prototype. Its final version is mostly the same as the one described back in section 6.2.3.

10.2.1 Structure

The psychological literature on goals and motivations (see section 3.3.1) focuses on long-term, high-level life goals rather than concrete, short-term goals that could be related to communicative behavior. The examined taxonomies roughly align with the Big Five personality model or the Interpersonal Circumplex, indicating that a connection to communicative goals exists. However,

this provided little information about what exactly those communicative goals should be.

In contrast, research in human-agent interaction tends to link behaviors directly to personality traits without reasoning about the underlying goals (see section 5.3). Success in conveying a specific personality is generally measured subjectively, but those subjective measures are less appropriate for a decision-theoretic approach.

Those works in human-agent interaction that do focus on goals - or rather, specific *intentions* - tend to cover pragmatic, functional goals (see section 4.3). When personality is related to goals, it is mostly done in the context of event appraisal and simulating emotional responses [44, 103].

Since there was no direct reference for defining turn-taking goals and desirable outcomes, goals were derived from the established definitions of the different personality traits. Furthermore, to model both the goals' functional relevance and that rooted in personality, several weighting factors were used to represent the goals' activation that is proposed in the OCC2 model [98, p. 55].

Attention Model

One core idea for this model was that gaze behavior is rooted in attention management. As detailed in section 3.4, humans turn their eyes toward relevant objects or people whose communicative signals they intend to observe. However, when they focus on planning or delivering their own contribution, they tend to avoid looking at the interaction partner. Previous research by Mehlmann et al. [85, 83], Skantze et al. [122], or Andrist et al. [4] showed that these gaze patterns can be transferred to human-robot interaction and give humans the impression that similar processes take place inside the robot's "mind".

First Iteration An early draft for the influence diagram (see figure 10.1) linked the agent's gaze direction to the chance of obtaining visual information, which in turn contributed positively to the goal "get information" but negatively to the goal "avoid overload". These two goals depended on the current progress of either participant's verbal contribution, reflecting the following assumptions.

- Cognitive load is high when the agent is in the planning phase but low otherwise.
- Feedback from the listener is needed while transmitting the MNI but less so after that point. It is not needed outside speaking.

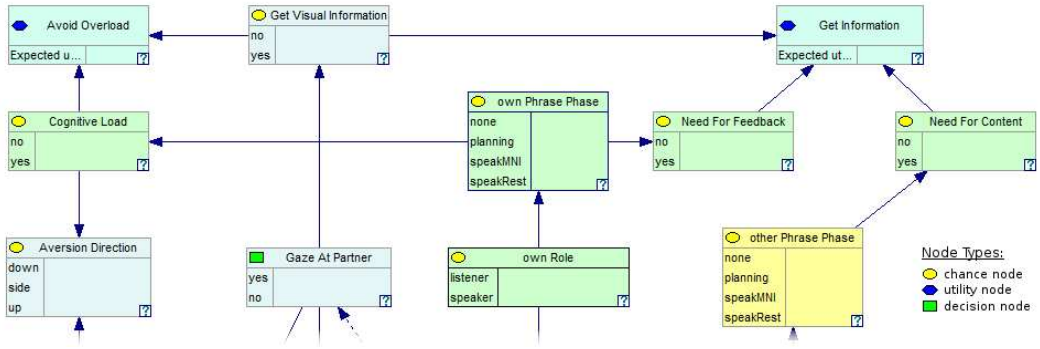


Figure 10.1: Excerpt of an early draft for the influence diagram of the interactive prototype. The green nodes represent the agent's own conversational state. The blue nodes represent its reasoning about its goals ("get information" and "avoid overload") and available actions ("gaze at partner" or don't). Yellow nodes represent the agent's belief about its interaction partner. The icons in each node's upper left corner indicate the node type.

- Content from the other participant is needed while the latter is transmitting the MNI, but not while they are listening. Otherwise, it can be either.

Second Iteration However, this structure offered no straightforward way to model the influence of the agent's personality on its gaze. As mentioned in section 3.4.3, humans associate higher amounts of gaze with closer and more positive relationships, as well as higher dominance. The interpersonal attitude, in turn, is related to personality traits like Extraversion and Agreeableness (see sections 3.2.3 and 3.2.4).

The network was therefore restructured to explicitly model two intermediate states: *Attention towards the Self* and *Attention towards the Other*. The former depends directly on the progress of the agent's verbal contribution. In contrast, the latter depends on the agent's objective *Need for Feedback* and its subjective *Interest in the Other*. Said interest, in turn, is derived from both the *Curiosity* rooted in the agent's *Openness* and the *Relationship Intensity* apparent from the expressed *Affiliation*. The relevant excerpt of the network is shown in figure 10.2.

Two more nodes were inserted to make the mapping onto the gaze direction easier to understand. One was the agent's *Cognitive Load*, and the other was the *Cognitive Target*. The idea behind this separation was that the latter would distinguish between seeking the other participant's gaze and avoiding it to focus on the agent's inner processing. At the same time, the level of the cognitive load was to determine the vertical direction, approximating the

observations in related work. As a rule of thumb, gazing up is perceived as being deep in thought [4, 122] while gazing down is perceived as merely holding the turn [4].

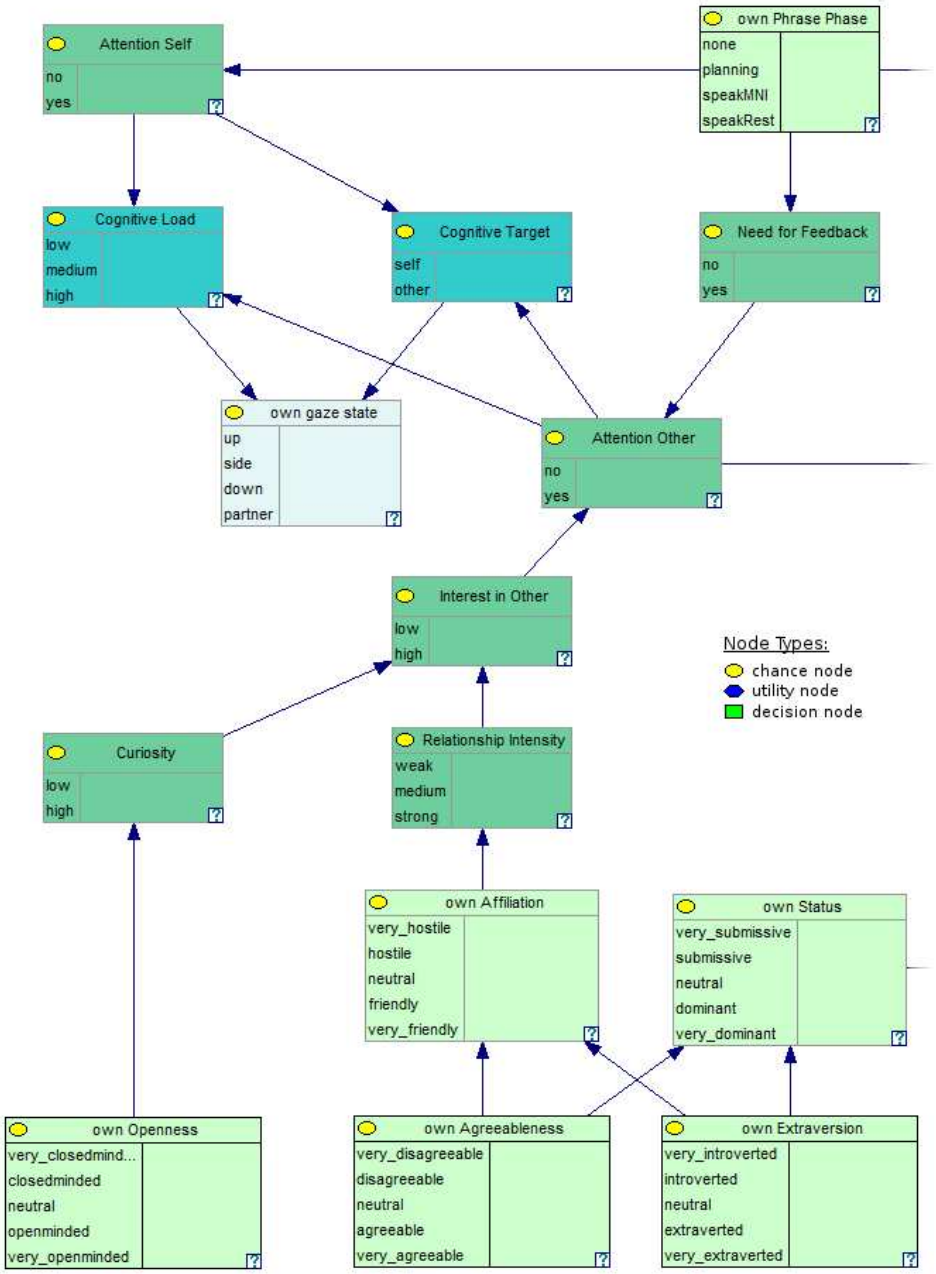


Figure 10.2: Excerpt of a later draft for the influence diagram of the interactive prototype. The green nodes represent the agent's own personality configuration and conversational state. The blue nodes represent its reasoning about its cognitive state (load and target) and available actions (gaze direction).

Third Iteration The attention model was refined in the influence diagram's final version, as shown in figure 10.3. The nodes *Attention towards the Self* and *Attention towards the Other* were first merged into one node named *Attention Target* that was eventually split into separate nodes for each modality. In the presented setup, these are the agent's *own visual attention* and *own speech attention*. This change was made because different mutually exclusive attention targets were more plausible than dividing attention between two targets. At the same time, gaze and speech behaviors were observed to be independent to some extent.

The *Cognitive Load* node was replaced with one representing the agent's *own feedback need*. Instead of being derived from the cognitive load, the gaze direction is now derived from both the utterance progress and the attention target. At the same time, the node *own feedback need* is used to activate the goals of seeing and hearing.

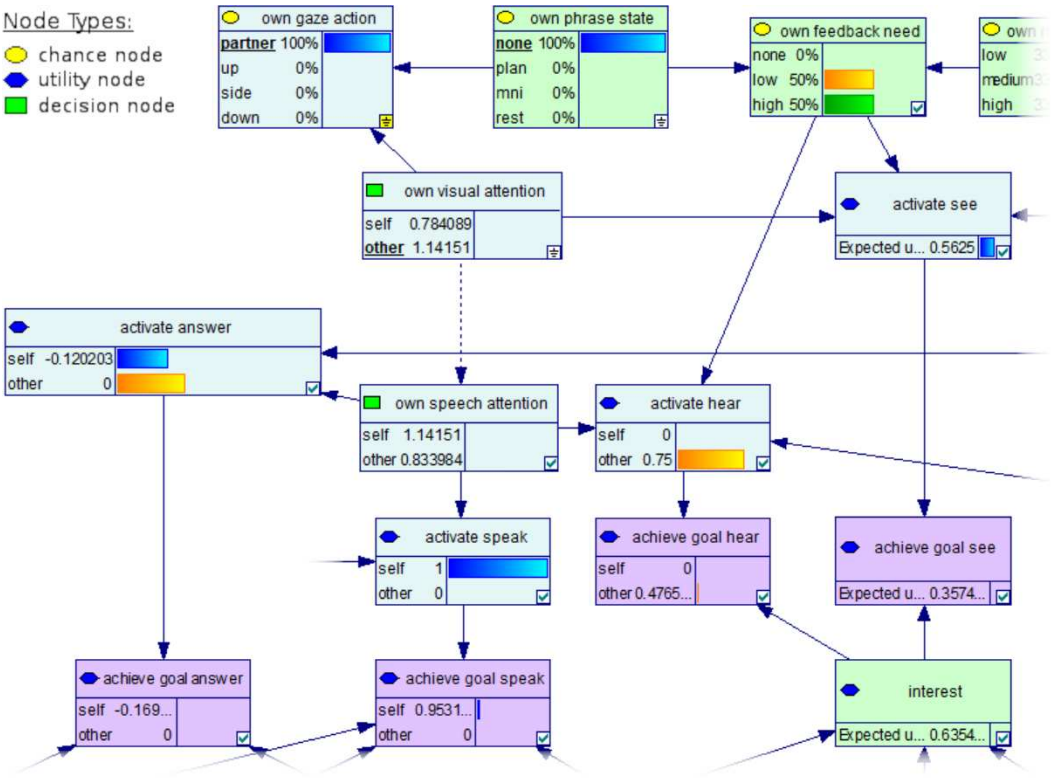


Figure 10.3: Excerpt of the final influence diagram, showing the decision nodes "own speech attention" and "own visual attention" for choosing the respective attention target.

Goals Definitions

In this version of the influence diagram, an interaction goal is represented by three or more separate nodes. This was necessary for clearly distinguishing between the objectively measurable prospect, its contribution to achieving the goal in question, and the relevance that the goal has for this particular agent.

Prospects Proper reasoning about costs and benefits requires clearly defined consequences that are composed of objectively measurable effects. However, such measurable effects are hard to find in psychological literature.

This is partially linked to the problem of abstract goal definitions. If goals are defined on the level of "be competent" or "be liked", there is no direct measure of success. In fact, the degree to which certain behaviors elicit perceptions of competence or likability is traditionally measured in an inherently subjective way. Many studies use either self-report questionnaires or the judgment of external observers to measure these constructs. In the first case, people may be tempted to rate themselves more favorably. As for the second case, Argyle and Little noted that the apparent personality of a person depends on context factors such as their role and the presence of certain observers [11].

Abbas advises against subjective ratings and recommends measures of value that are grounded in concrete attributes such as monetary costs or physical properties [1, p. 162-167]. Applied to the problem of conversational timing, this means that the abstract "degree of control" (see chapter 9) would be better represented by, for instance, the number of seconds spent speaking or the number of messages that were successfully transmitted.

Concrete Short-Term Goals Consequently, the agent's goals had to be defined on a very low level of abstraction. The events that would lead to their achievement had to be objective and deterministic given the world state, with all uncertainty or subjective interpretation limited to the observation of the world state itself.

Following the definition of *active pursuit goals* in the OCC2 model [98, p. 50], the most basic goals that the agent could have is to perceive the other participant and to speak its own verbal contribution. These directly affect the agent's behavior, and so they are modeled explicitly. Higher-level goals are only modeled indirectly, with the understanding that an undesirable state implies a goal of changing that state for the better. Using the OCC2 terminology, a conversational agent has several long-term *interest goals*.

- The agent wants to see its own needs fulfilled.
- The agent wants to see a minimal delay for urgent actions.

- The agent wants to act in line with its personality traits.

These interest goals contribute to activating the concrete action goals, which in turn increases the matching behaviors' utilities. Mathematically, the agent's interests are mapped to weighting factors that are explained in the following.

Personality-based Goal Relevance

In the context of floor management, overlaps and delays are two observations that have been linked to perceptions of personality or interpersonal attitude (see section 3.4). Delays are intuitively tied to wasting time, indicating a lack of discipline and, therefore, low conscientiousness. Alternatively, they can be interpreted as being hesitant and thus showing lower self-confidence and assertiveness. The latter is supported by several studies that showed delays to appear less dominant and less extraverted [130, 27, 47, 64].

Similar facets were identified for the other personality traits. Based on the definitions of those traits in psychological literature, several associated interests or acting tendencies were selected to serve as weighting factors. Figures 10.4 and 10.5 show how the agent's goals are connected to its personality traits and interpersonal attitude via those intermediate characteristics.

The *Conscientiousness* level translates to a value of *duty* in the range $[0.0, 1.0]$. This duty serves as a weight for two speech-related goals, *speak* and *answer*. The first one is the agent's intrinsic motivation to transmit the current message, while the second one represents the motivation to provide the information that the other party seeks.

Seeking to fulfill someone else's goals implies a care for and psychological closeness to the other party. It can also be connected to the *positive face* [20] and the *Affiliation* component of the Interpersonal Circumplex. (See section 3.4.3 for details.) Consequently, another weighting factor for the goal *answer* is the *care for the other* that is derived from the agent's *Affiliation* level.

Its *Status*, in contrast, does not influence the affiliative goal. Instead, it is mapped to a value of *assertiveness* that contributes to the relevance of the *speak* goal. Unlike the agent's sense of *duty*, this represents the idea of an interlocutor speaking mainly to occupy the conversational floor and not to advance the conversation.

Another factor that adds weight to the *speak* goal is the character's *agitation*. It is derived from the character's *Neuroticism* and expresses the urge to break the silence, for example out of impatience or excitement. Besides *agitation*, this personality trait is also translated to the factor of *watchfulness* that simulates the phenomenon of looking out for signs of danger or the other's opinion of oneself.

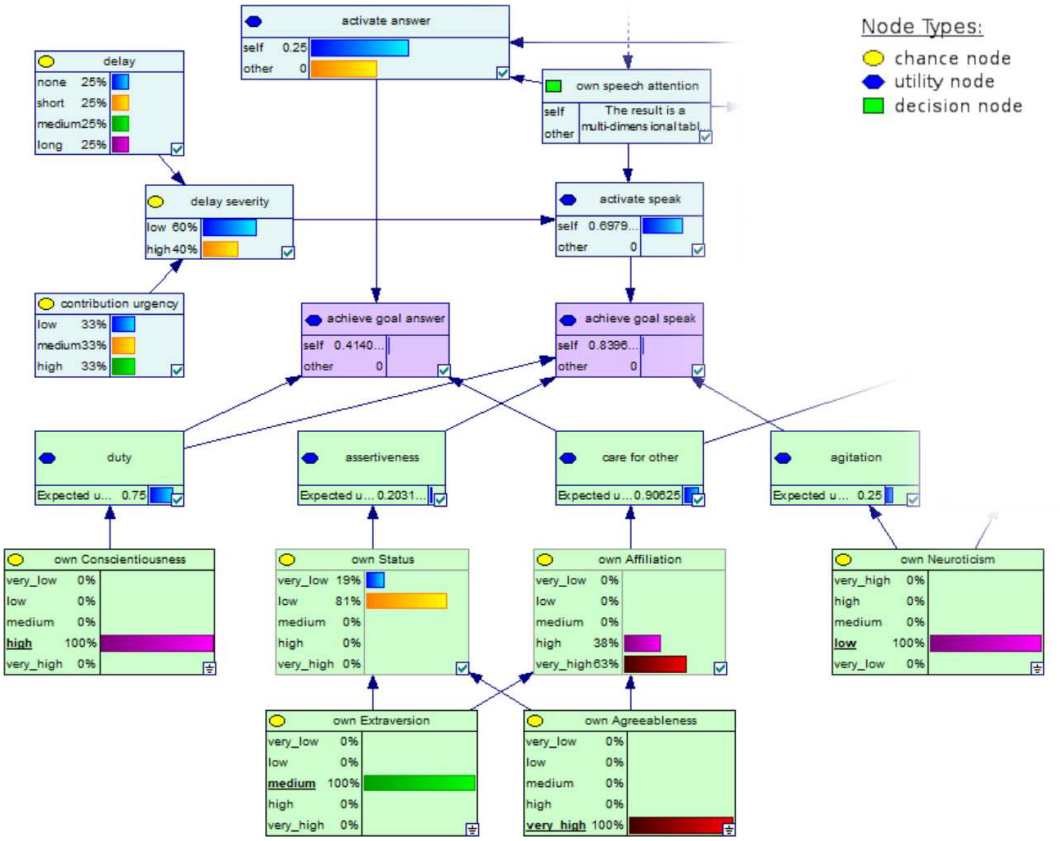


Figure 10.4: An excerpt of the final influence diagram, showing the speaking goals and their connection to the agent's personality traits.

Watchfulness and the aforementioned *care for the other* are averaged to form an overall *interest* factor, along with a third factor named *curiosity* that represents the level of the agent's *Openness*. This *interest* is then used as a weight for the goals *listen* and *see*. This relationship is inspired by the conflicting findings about what makes a human look at another. In other words, the generic *interest* hides the possible reasons for why an interlocutor might show interest in the other and focuses on the ways in which it is expressed.

Functional Goal Activation

Personality traits are not the only factors that bring a goal into or out of focus. Some goals are irrelevant in certain dialogue phases, or impossible to achieve when the agent chooses a specific action.

Figure 10.6 shows an excerpt of the influence diagram, focusing on the nodes that represent the goals "speak" and "hear" as well as the neighboring nodes that contribute to these goals. The nodes labeled "achieve goal X" are multiattribute utility nodes that multiply the personality-specific relevance of

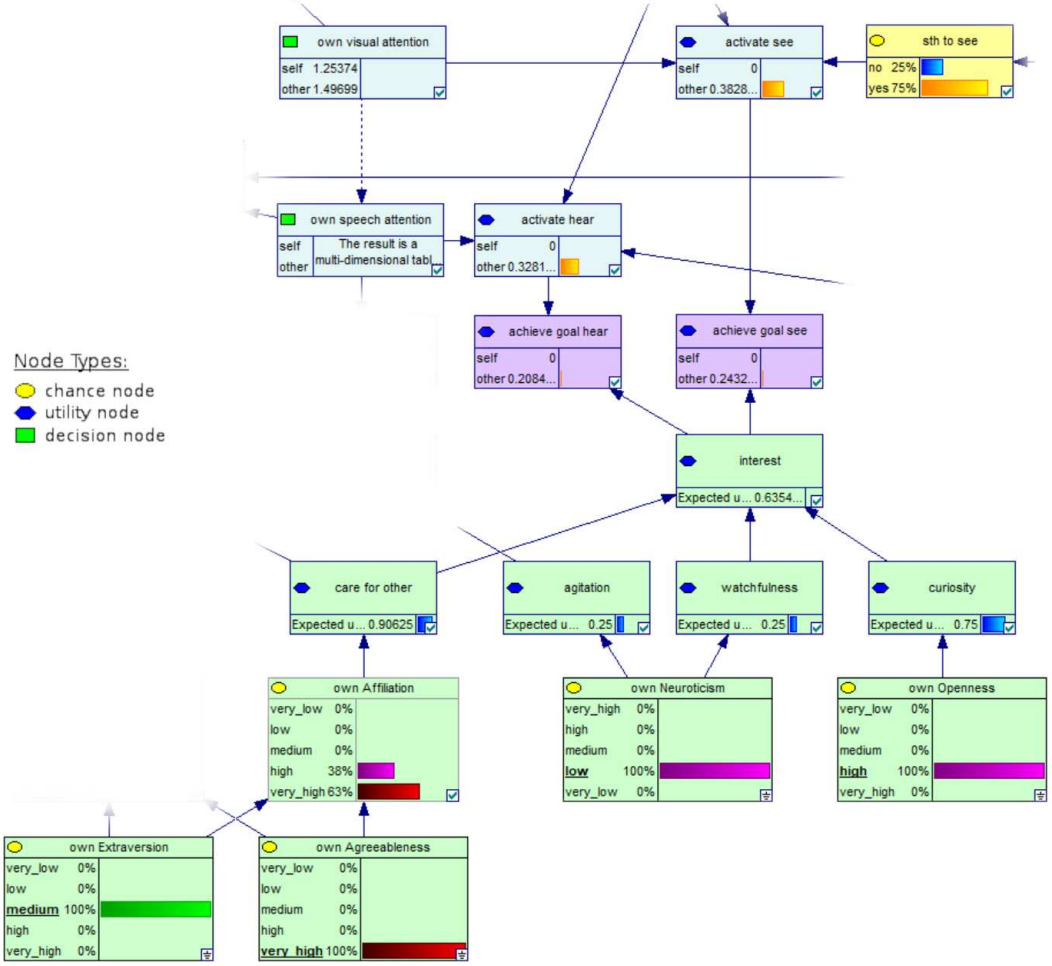


Figure 10.5: An excerpt of the final influence diagram, showing the observation goals and their connection to the agent's personality traits.

the goal with the degree to which it is activated.

Here, "activated" means "objectively relevant in the current situation *and* can be achieved by the selected action". In other words, if the agent's verbal attention target is the other participant, it cannot utter its own contribution and the goal "speak" is forcefully disabled by multiplying the contextual activation by zero. Consequently, the unachievable goal does not contribute to the total expected utility of selecting that attention target. More details on that will be provided in section 10.2.3.

10.2.2 Probability Distributions

The parameters of the first influence diagram were manually derived from theories and study results found in the literature. Trying to apply the same methods for modeling gaze behavior revealed that there was a lack of concrete,

numerical parameters for nonverbal behavior in psychology. Those parameters that could be found (for example, in those studies summarized by Argyle and Cook [8]) usually referred to the interaction as a whole rather than specific moments. Those studies also tended to look at one specific aspect, such as perceived dominance or liking, in isolation from speaking turns. Consequently, the parameters that are used in this thesis were taken from related works in computer science.

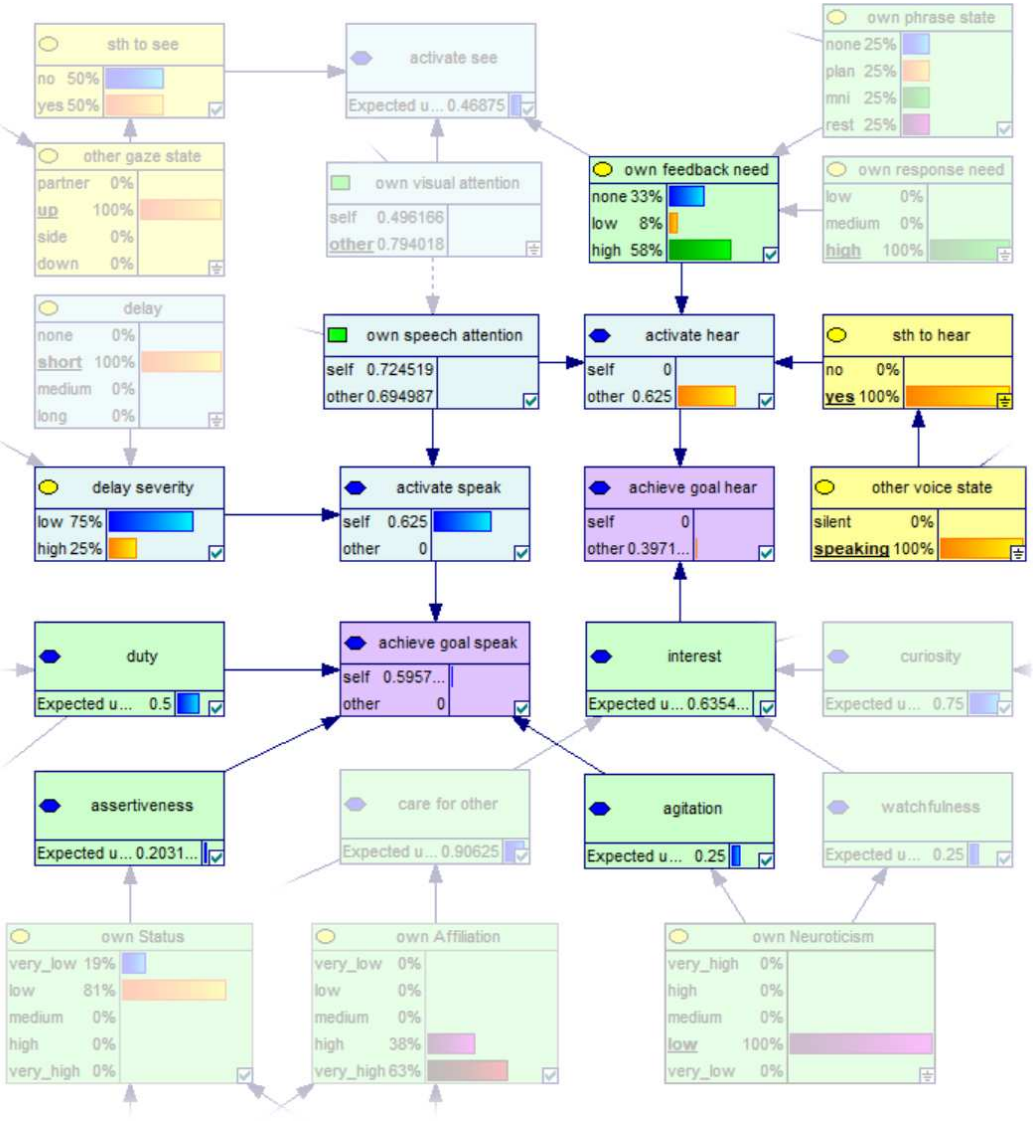


Figure 10.6: Excerpt of the final model, showing the factors for weighting and activating the two goals "speak" and "hear".

Gaze

Probabilities for the gaze direction were taken from computer science, specifically from Andrist et al. [4]. The state "partner" signifies gazing at the interlocutor and is observed only when the network decides that the other party is the better target for the agent's visual attention. When the target is the self, the probability distribution instead covers the remaining three states "up", "side", and "down" with the percentages that Andrist et al. had observed in their video recordings of human-human conversation. (See also section 6.3.1 for the exact numbers.)

Delay Severity

Ideally, the actual duration of the *delay* would be represented as a concrete prospect. However, these may be different between cultures, and it is not known whether humans actually pay attention to the precise time or even if the severity scales in a linear manner. Therefore, the duration was discretized to "none", "short", "medium", and "long". The precise intervals are configured outside the behavior model, and the agent uses them to label the tracked duration when synchronizing its *situation parameters* with the Bayesian network.

The same considerations were applied to the inherent *urgency* of a specific dialogue act. The agent uses manually defined numerical values in the range $[0.0; 1.0]$ for setting this situation parameter and maps that value to the labels "low", "medium", and "high" before setting the observation in the Bayesian network.

Both *delay* and *urgency* are uniformly distributed a priori, but this becomes irrelevant as soon as the agent starts tracking the conversation context variables. They combine into the total *delay severity* as described in section 6.3.4.

Interpersonal Attitude

The probability for observing a certain *Status* respectively *Affiliation* level was calculated the same way as it was done for the non-interactive prototype. (See section 6.3.2 for details.) The only difference is that the rotation angle was slightly adjusted to $-37.964\,49°$. As explained in section 9.4.3, this angle minimized the error between the predicted *Status* value and the ratings given by study participants. Of all angles with that property, it was also closest to the theoretical value of $-37.5°$ that was based on psychological literature [80, 79, 39].

10.2.3 Utilities

As explained in section 6.2.3, an interaction goal in this influence diagram consists of at least three different nodes. A multi-attribute utility (MAU) node combines the personality-based relevance of the goal with its context-dependent activation, both of which are either basic utility nodes or more MAU nodes.

Goal Activation

The activation of $goal_i$, as mentioned before, is composed of the *relevance* given the observed interaction *context* (including such variables as the utterance progress or the urgency of the agent's contribution) and the *possibility* of achieving the goal with the attention set to $target_j$.

$$relevance(context, goal_i) = \sum_{l=1}^{\#factors(goal_i)} utility_l(context)$$

$$possibility(goal_i, target_j) \in [0.0, 1.0]$$

$$activation_i(context, target_j) = possibility(goal_i, target_j) * relevance(context, goal_i)$$

In the final version of the influence diagram, the *delay severity* is discretized into the two levels "low" and "high". The first one activates the goal "speak" halfway, whereas the second one activates it fully. Achieving the goal is fully possible (1.0) if the verbal attention target is the self, but impossible (0.0) otherwise.

The goal "hear" is activated based on the agent's need for feedback. A "high" feedback need leads to a full activation, a "low" one to a halfway activation, and the level "none" deactivates the goal. The activation is moderated by both the verbal attention target and the speech state of the other participant. The goal is achievable (1.0) when the verbal attention target is the other and impossible to achieve (0.0) otherwise. It is also considered fully possible (1.0) when the other participant is speaking. When they are not, the goal is still partially attainable (0.5) because the absence of information can also be revealing. Both *possibility* values are multiplied to obtain the total moderation factor.

The goal "answer" is defined very similarly, except that the activation depends on the assumed feedback need of the other participant. In this case, the verbal attention target equals the speech state because when the influence diagram is queried for its decision, the agent knows that it will speak when

permitted. Consequently, the *possibility* factor is 1.0 for target "self" and 0.5 for target "other".

Goal Achievement

A goal can only be achieved if it is activated, and the achievement only matters if the agent's personality makes it care about that goal in the first place. Consequently, the goal's activation and relevance are multiplied.

As explained in section 10.2.1, every level of the interpersonal attitude dimensions and the personality traits is mapped to one or more weight factors that are represented by a basic utility node. In this version of the influence diagram, this is a linear relationship with the lowest level of the trait mapped to 0.0 and the highest level mapped to 1.0.

For each $goal_i$, the utility of selecting a given attention target is calculated as follows.

$$utility_i(observation, target_j) = (activation_i(observation, target_j)) * (\sum_{k=1}^{\#weights} weight_{i,k})$$

The expected utility of selecting a given attention target is then calculated as the sum of the goal utilities.

$$EU(context, target_j) = \sum_{i=1}^{\#goals} utility_i(context, target_j)$$

10.3 Implementation

Compared to the non-interactive prototype, the implementation changed massively. The knowledge management was overhauled, and a framework was created to connect all participants, from humans to static and adaptive agents, in a structured way. Furthermore, true sensor input for both speech and gaze was required to enable a natural conversation.

10.3.1 Architecture

Figure 10.7 shows the main components of the interactive setup. An incremental speech recognizer processes the user's voice, and the message content is stored in the agent's memory for use within the dialogue manager. At the same time, their raw voice activity and current gaze target are set as observations in the influence diagram. They inform its decisions about the agent's

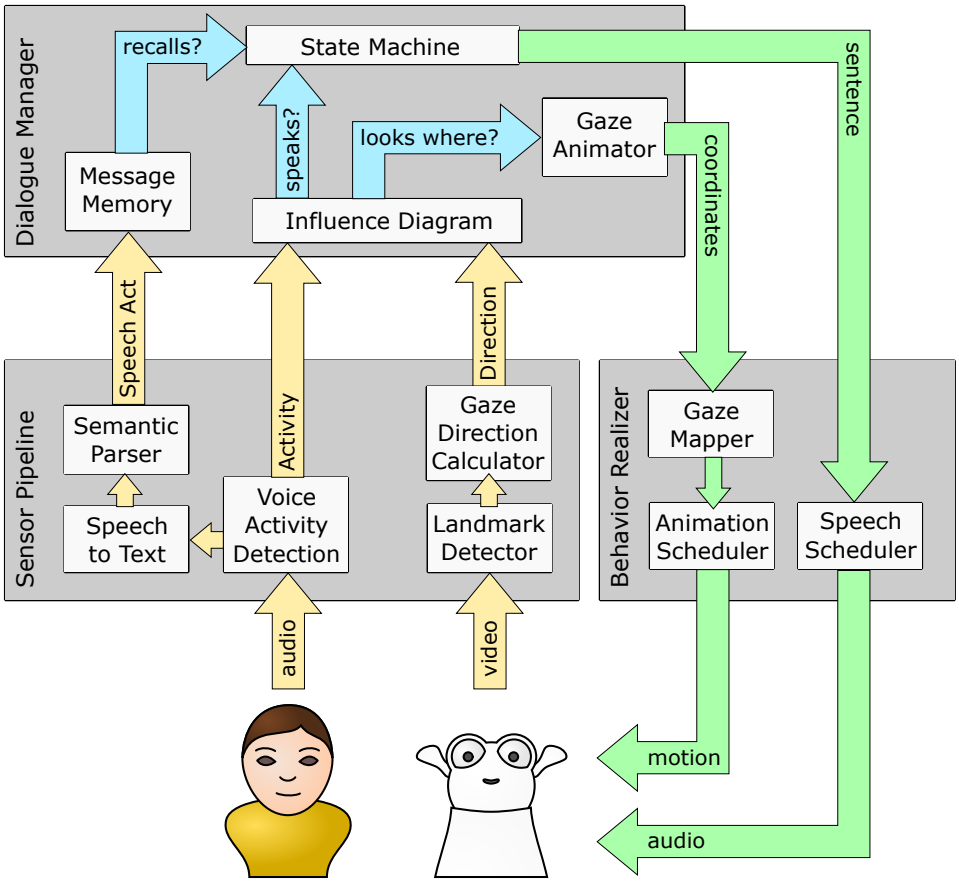


Figure 10.7: Architecture of the interactive setup.

visual and verbal attention, which then regulate the dialogue flow and animate the agent's own gaze. Finally, the resulting speech and gaze commands are sent to the robot's behavior realizer.

Participant Framework

All data concerning the interaction's participants is managed by the *Participant Framework* that was presented in chapter 7. This makes it easier to set up different constellations of humans, virtual agents, and social robots or to switch from non-adaptive agents to learning agents later.

Both human and computer-controlled *participants* are connected to a message hub for exchanging information in various modalities. *Agent participants* additionally have a semantic memory for storing the exchanged messages and a pool of situation variables of which they keep track. This basic agent type can then be extended with an influence diagram, an implementation for reinforcement learning, or both.

Dialogue Management

The **InterruptibleExecutor** class for Visual SceneMaker (see 9.3.2) was restructured to make the dialogue setup as modular as possible. The interactive prototype no longer uses dedicated Executors to handle the shared information or the Bayesian network. Instead, everything related to knowledge management and decision-theoretic reasoning was moved to the **InterruptibleAgentParticipant** class and its members.

10.3.2 Knowledge Management

The non-interactive prototype used a simplified knowledge representation because the focus was on the overlap of speech activity rather than the content of the overlapping utterances. However, study participants remarked on the lack of semantic context that made it hard to distinguish between domineering or enthusiastic behavior (see section 9.4.3).

In contrast, semantic information was going to play a greater role in the interactive scenario. Therefore, new classes were created to represent the feature structures and retrieve messages that contain the specified data.

Data Structure

The data structure for the exchanged messages is based on the Dialogue Act Markup Language (DiAML) [23, 24], as well as the typed feature structures used by Mehlmann et al. [83, 84]. For exchanging and comparing these feature structures, they are represented in the JavaScript Object Notation format (JSON)[1]. Every message holds the following properties:

- **source:** the name of the agent sending the message
- **modality:** the modality of the message *voice*, *speech* or *gaze*
- **time:** the time at which the message is perceived
- **act:** the communicative act transmitted in this message

Following the definitions in section 2.4.2, the act contains two elements:

- **function:** the communicative function, such as "inform", "request" or "social"
- **content:** the semantic content, which in turn holds an arbitrary set of key-value pairs

[1] https://www.json.org/

```
{
    "*type*": "Message",
    "sender": "user",
    "addressee": "any",
    "modality": "speech",
    "act": {
        "function": "social",
        "content": {
            "type": "greeting"
        }
    }
}

{
    "*type*": "Message",
    "sender": "user",
    "addressee": "any",
    "modality": "voice",
    "act": {
        "function": "inform",
        "content": {
            "state": "speaking"
        }
    }
}

{
    "*type*": "Message",
    "sender": "user",
    "addressee": "any",
    "modality": "speech",
    "act": {
        "function": "request",
        "content": {
            "info": "price"
        }
    }
}

{
    "*type*": "Message",
    "sender": "user",
    "addressee": "any",
    "modality": "gaze",
    "act": {
        "function": "inform",
        "content": {
            "state": "partner"
        }
    }
}
```


Figure 10.8: Standardized message format for transmitting information between participants.

The same two elements form the MNI and are given as the parameters of the *minInfo* action marker. Figure 10.8 gives some examples of these feature structures.

Semantic Memory

Every agent stores two types of information: the exchanged messages and the current state of the world around it.

The **CommunicationMemory** stores the messages that the agent received. It also provides a method for querying if a message with the given data is stored.

- **source:** the sender of the message, who is either the human user or the agent itself
- **modality:** the modality over which the message was transmitted
- **act:** a communicative act to be compared with that of the stored messages
- **keep:** a flag indicating whether matching messages should be removed from the memory

For a message to match the query, all parameters with values other than "unknown" or "any" must be present. The function of the act must be identical, and the content must contain at least the given key-value pairs. The queries themselves are attached to conditional edges between Visual SceneMaker's state nodes.

The **Situation** holds a collection of **SituationParameter** objects. Their values are synchronized with the Bayesian network at appropriate times. Some of them are used to configure the agent, for example, to set the level of *Extraversion* or to distinguish between long and short amounts of *delay* since the last speaking attempt. Others increase or decrease automatically, either over time or upon triggering events. Here, it is used to track the delay duration[2]. The Bayesian network is updated on the following occasions:

- The dialogue advances to a new utterance.
- The *phrase state* is updated.
  - The speech command is sent to the agent's behavior realizer.
  - The agent's TTS engine starts producing audio.
  - The TTS output reaches the end of the MNI.
  - The agent's TTS engine stops producing audio.
- The agent finds a given speech act in its memory and sets the associated response need for the speaker.
- The agent observes voice activity from the other party.
- The agent observes gaze activity from the other party.

Messaging

A central *MessageHub* ensures that messages sent out from one participant reach all other parties. It receives standardized messages (see section 10.3.2) from either a computer-controlled agent or an external sensor pipeline. If the addressee is specified, the message is forwarded directly to the agent in question. Otherwise, it is broadcast to all agent participants. Since human participants can observe the actions of a robotic or virtual character, the hub does not forward any messages to them[3].

[2] Automatically changing parameters were originally implemented for tracking engagement in a learning agent. Another potential use could be a simple simulation of short-term affective states, such as disappointment in case of an interruption

[3] A potential exception could be a remote Wizard-of-Oz experiment interface that does not allow for direct observations.

Agents send out *Message* objects whenever an utterance arrives at the end of the MNI. Upon receiving one from the hub, it stores it in its local memory and, if necessary, uses it to update its own knowledge about the interaction context. In the final prototype for this thesis, voice activity from any participant is used to reset the tracked *delay* while the other participant's gaze target is directly set as an observation in the Bayesian network.

The dialogue manager can remove messages from an agent's memory, either while querying it for a matching speech act or by clearing it completely.

10.3.3 Procedural Gaze Animation

Adding the modality of gaze was a major change compared to the non-interactive prototype. The gaze animation was coded into the *InterruptibleAgentParticipant* class to minimize performance issues. Experience had shown that modeling the agent's gaze as a parallel section of the state machine introduced delays that strongly interfered with the fine-grained timing necessary for natural gaze patterns.

Animation Scheduling

Two duration parameters were used to control the animation. The *shift duration* parameter defined the number of milliseconds during which the agent's gaze moved from the old target position to the new one. The *fixation duration* parameter defined the number of milliseconds that the agent's gaze would linger on the target after reaching it.

The two durations were added up to form the interval at which the scheduling thread selected a gaze target and sent the appropriate command to the agent.

Target Selection

As explained in section 6.2.2, each modality that the agent uses is linked to an attention target. The decision node *own visual attention* influences a chance node named *own gaze action* that specifies the gaze direction. (See section 6.3.1 for the conditional probability distribution at this node.)

Whenever the animation thread updates the agent's gaze, a gaze target label is randomly drawn according to the active probability distribution at *own gaze action*. The targets themselves were defined in a configuration file that associated each label with a set of local coordinates. The following four targets (also shown in 10.9) were defined for a face-to-face dyadic setup.

- **partner:** looking straight ahead
- **avert side:** looking to the side, but on the same level
- **avert up:** looking up and slightly to the side
- **avert down:** looking down and slightly to the side

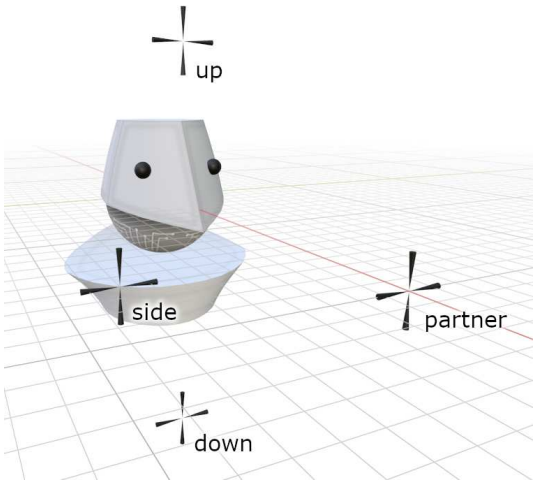


Figure 10.9: Gaze targets relative to the agent's head.

Interpolation

While the selected gaze targets reflected the probability distribution, the resulting behavior appeared very unnatural due to the sudden changes in movement direction. Therefore, a smoothing approach was needed.

The implemented solution is inspired by the representation of the emotion layer in Gebhard's ALMA [44]. The implementation of ALMA uses three important concepts for reconciling conflicting emotions.

- **Activation:** Whenever an emotion is triggered, its position in the PAD space (see 3.2.1) gains a certain amount of weight.
- **Decay:** As time passes, every emotion's weight is gradually reduced until it returns to 0.
- **Center of Mass:** The weighted average of the emotions' PAD coordinates represents the current emotional state, for example, when calculating its influence on the mood layer.

In the work presented here, the gaze targets were activated in a similar way. Each update step began with reducing the weights of all targets by the numerical value defined in the *activation decay* parameter. A new target was then drawn randomly, and the value of the *activation increase* parameter was added to its weight.

The gaze target $\vec{g}_{avg}$ was calculated as follows, with n being the number of predefined gaze targets $\vec{g}_i$ and w_i being their weights.

$$\vec{g}_{avg} = \frac{\sum_{i=1}^{n} w_i \vec{g}_i}{n}$$

The resulting 3D coordinates were then sent to the agent's behavior realizer. Consequently, repeated activation of conflicting gaze targets (such as *avert up* and *avert side*) caused the agent to turn its head towards an intermediate point instead of alternating between the two orientations. When one gaze target was drawn more frequently than others, the agent's visual focus gradually shifted towards this one while the remaining targets lost influence.

10.3.4 User Input Recognition

To understand the user's turn-taking intention, their nonverbal signals need to be detected in real-time and discretized to the labels that the influence diagram can handle. Furthermore, the MNI of their verbal contribution needs to be identified so that the agent can interrupt under the right circumstances.

Gaze

MediaPipe[4] is used to recognize facial landmarks from a video stream. Said video stream could be obtained from, for example, a laptop's inbuilt webcam or from a robot's eye camera.

In the next step, the direction of the face and eyes are calculated. Four landmarks on the face outline were selected to obtain the vertical and horizontal axis. Their cross-product was calculated, normalized, and taken as the direction of the face as a whole.

Calculating the eye direction from just the iris landmarks proved to be too noisy in practice. Consequently, the direction vector was determined with the help of an additional reference point. For this, the point between the inner eye corners is shifted back by a few centimeters along the face direction vector. The center between the outer iris landmarks is then used to calculate the gaze vector.

[4] https://google.github.io/mediapipe/

Both the direction of the face and eye are discretized based on angle thresholds. In the prototype, those angles are ±10° for the face and ±5° for the eyes. Finally, the eye gaze direction is mapped to one of the four state labels used in the influence diagram.

Speech

An incremental speech recognizer was required to detect the end of the MNI in the user's speech. Specifically, the MNI is a speech act that is put out as an intermediate NLU hypothesis and allows the dialogue manager to select a valid response. The setup presented here uses the Retico framework[5] by Michael [88].

Retico provides a module for offline speech recognition using wav2vec[6] [13]. This module is used to transcribe the audio before extracting its meaning.

Since the transcription may contain incorrect spellings or similar errors, it was unsuitable for parsing with a context-free grammar or regular expressions. Instead, the text is processed by Rasa Open Source[7]. While there is an official Rasa module for Retico, the one used here was built from scratch so that it would be compatible with the rest of the presented setup. For instance, the official Rasa module expects the output of the module for Google's speech recognition, which is very different from that provided by the wav2vec module.

It was also hard to understand how complex feature structures could be attached to Rasa's training data, so the "RasaModule" in this prototype maps the unique intent identifiers to the speech act that the dialogue manager expects. The mapping for an extension of the "Salesperson" scenario is given in table 10.1.

Voice Activity

While the wav2vec module already contains a detector for voice activity, this activity is not accessible from outside the module. Therefore, an additional custom module was implemented that provides the user's voice activity for the agent's behavior reasoning.

The "VoiceActivityDetectorModule" was implemented based on WebRTC-vad[8], a Python module that is already a requirement for Retico. In fact, the same voice activity detection is also used by the wav2vec module. The labels for the module's output - "speaking" for activity, "silent" otherwise - are given as arguments that can be easily changed if needed.

[5] https://github.com/retico-team
[6] https://github.com/retico-team/retico-wav2vecasr
[7] https://www.rasa.com/
[8] https://pypi.org/project/webrtcvad/

Rasa intent	function	content
greet	social	{"type":"greeting"}
goodbye	social	{"type":"goodbye"}
has_vacuum	inform	{"subject": "user", "property": {"has_vacuum": "yes"}}
buy	accept	{"offer": "vacuum"}
accept	accept	{}
decline	reject	{}
ask_colors	request	{"info": "colors"}
ask_technology	request	{"info": "technology"}
ask_price	request	{"info": "price"}
complain_price	inform	{"property": {"price":"expensive"}}
accept_price	inform	{"property": {"price":"fair"}}
unsure	inform	{"subject":"user", "property": {"certainty":"low"}}
not_interested	inform	{"subject":"user", "property": {"interested":"no"}}

Table 10.1: The mapping between the intent provided by the Rasa module and the communicative act that the dialogue manager will look for.

A helper module named "MessageCreatorModule" then wraps the state label in a message object that is transmitted to the dialogue application.

10.4 Evaluation

Several sample interactions were recorded to confirm that the behavior model produced plausible and distinct behavior patterns. Those took place between a human and an autonomous agent that was controlled by the presented behavior model. The recordings were then compared to see where the generated behavior differed between the respective personality configurations.

10.4.1 Scenario

First of all, a scenario had to be defined to showcase and compare the different agent personalities. It was then implemented for a human-agent dyad that would provide the sample recordings for analysis.

Conversation Topic

One major challenge was finding a suitable conversational topic for showcasing the behavior. It needs to invite interruptions but also be a plausible conversation that could happen between two humans. Unfortunately, the combination of these requirements quickly leads to a large and complex domain.

Turn-taking conflicts arise most likely when one participant is holding the floor for a long time or delays the conversation by planning their contribution thoroughly. Consequently, the conversation topic should invite complex replies and commentary from both participants. A scenario's usefulness for turn-taking hinges on the agent's ability to give meaningful opinions and advice rather than short, generic acknowledgments. This, in turn, calls for sophisticated NLU capabilities and an equally sophisticated knowledge representation. The alternative is to limit what users can say about their day. While such limitations may be useful for experimental setups, they would make the scenario unattractive for live demonstrations.

Initial tests with a "how was your day" scenario (as used by, for example, Crook et al. [36]) were unsatisfactory due to the difficult balance between flexibility and predictability. Consequently, the salesperson scenario from the non-interactive prototype (see section 9.4.2) was revisited instead. It was extended with a branching dialogue flow and optional sections that provide more details about the topic. The chosen scenario has the following advantages:

- **Antagonistic relationship:** One side of the conversation, the resident, is unlikely to need a new vacuum cleaner. The salesperson, on the other hand, shows up uninvited and the former will probably want to send them away.

- **Long-winded presentation:** Even if the resident should be interested, the salesperson's product presentation consists of long sentences designed to make the interlocutor lose patience and interrupt.

- **Limited domain:** The conversation focuses on the vacuum cleaner and the attributes that might interest a customer. The salesperson also avoids small talk that could lead to unexpected answers.

- **Relatable topic:** Participating in this conversation requires little background knowledge or creative thinking. Users taking on the resident's role are most likely familiar with vacuum cleaners and people trying to sell them something.

Salesperson	**Resident**
Good day!	
I am from the company Dirt-B-Gone, and I would like to present our newest vacuum cleaner to you.	
Do you have a moment* to spare?	Uh, I don't know.
Believe me, you will not regret listening to me.	I already have* a vacuum cleaner.
	Alright.
I can assure you, compared to our new Slurp 380, your old vacuum cleaner will look like a stone age* relic.	Ah, I'm not so sure about that. My vacuum cleaner is still fairly new.
The Slurp 380 is the world's first vacuum cleaner with the revolutionary Piranhanado* technology!	With what technology?
The air duct design is inspired by the hydrodynamic properties of fish scales*.	
	Okay - how much does it cost?
Together with our patented turbine blades, this creates the strongest air flow* in the history of vacuum cleaning	Yeah, but what is the price?
It only costs 399* Euros.	Oh, that's very expensive!
And for only 99* Euros more, you can get 2 extra years* of warranty on top of that.	Uh, no. I'm not interested in that.
May I at least leave this brochure* with you?	No.
Alright*.	
	Goodbye.
Have a nice day!	

Table 10.2: Script for recording the interactive scenario.

Participants

The conversation took place in a dyadic setup between a computer-controlled agent and a human, specifically the author of this thesis. The former took on the role of the salesperson, while the latter played the resident.

Agent The initial recordings were done with a Robopec Reeti V1 that was connected to the dialogue manager using a wireless network connection. However, notable delays were observed in the robot's behavior, and it was hard to see whether they were intended by the turn-taking model or caused by the network communication.[9] Therefore, the robot was replaced by a virtual Klappmaul agent running on the same computer as the rest of the setup.

User To make the recordings comparable, the human's behavior had to remain as similar as possible between different conditions. Therefore, a script (shown in table 10.2) was prepared for this side of the conversation.

Adhering to the script was not entirely possible. On the one hand, human reaction times are hard to keep constant between sessions, and on the other, speech input was not always recognized correctly. Furthermore, gaze behavior was not scripted because it mostly happens subconsciously to begin with. Therefore, some slight variations occur in the samples.

10.4.2 Sample Interactions

To test the behavior model, the same conversation was recorded repeatedly with different agent personalities. Audio and video of both participants were captured for analysis, along with the states of the influence diagram.

Agent Personalities

Several personality configurations were prepared for the recordings. They were selected to cover different turn-taking styles but also represent believable archetypes for a salesperson. Table 10.3 shows the personality traits chosen for the final video recordings.

Recording

For each of the 4 salesperson archetypes, 7 sample interactions were recorded, resulting in 28 videos for analysis. OBS Studio[10] was used to capture the screen content. Figure 10.10 shows an example screenshot of the window layout.

[9] Later tests revealed that the delays were rooted in synchronization issues within the parallel state machines.

[10] https://obsproject.com

	aggressive	dutiful	friendly	lazy
Openness	low	medium	very high	medium
Conscientiousness	high	very high	medium	very low
Extraversion	very high	medium	high	medium
Agreeableness	very low	medium	very high	medium
Neuroticism	low	very low	medium	very low

Table 10.3: The personality traits for the different salesperson archetypes.

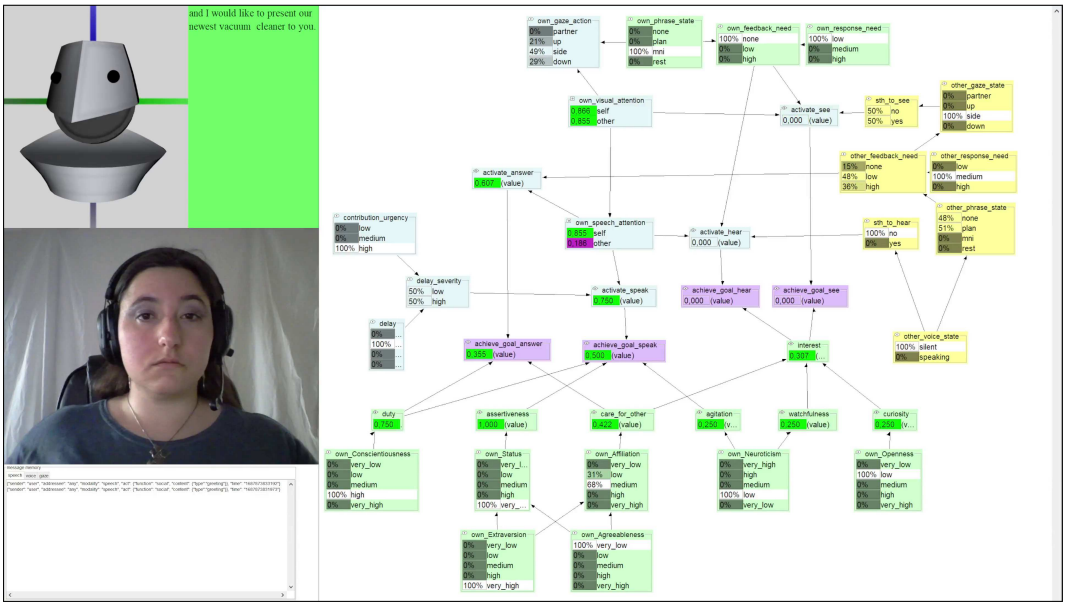


Figure 10.10: Screen capture layout for recording the sample sessions. *Upper left:* The Klappmaul agent and the utterance it is supposed to say. *Center left:* Video stream of the human interlocutor. *Lower left:* The semantic content of the verbal messages that the agent remembers. *Right:* The current state of the influence diagram.

- **Agent Behavior:** The Klappmaul agent was seen from the front, with a neutral gray background and an indication of the axes. The model with shoulders was chosen to emphasize the gaze behavior.
- **User Behavior:** The video stream from the laptop's camera was displayed in a dedicated window as part of the input processing pipeline. The audio signal was captured directly from the headset microphone.
- **Behavior Model:** The current state of the influence diagram was displayed in a dedicated window as part of the control application.
- **Semantic Information:** As soon as the dialogue manager proposed the next utterance, the corresponding text was displayed in a small window. A different window displayed the content of the agent's semantic memory, showing when the NLU module had parsed the user's speech.

10.4.3 Observations

In the first step, the videos were annotated with the "ELAN" tool [21][11]. The data obtained from the annotation was then exported to table format and analyzed in the "R" software environment[12]. The script written for analysis can be found in the appendix in section B.2.

participant	**M**	**SD**	min	max
user	1.475	0.808	0.288	4.315
agent	2.322	1.424	0.400	6.251
any	1.966	1.275	0.288	6.251

Table 10.4: Comparison of the utterance durations in seconds.

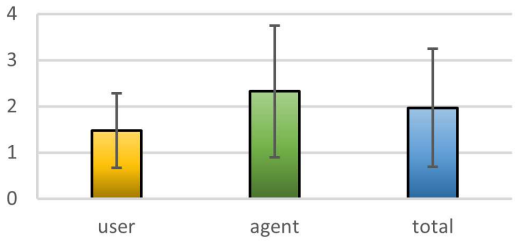


Figure 10.11: Comparison of the utterance durations in seconds.

[11] Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands, https://archive.mpi.nl/tla/elan

[12] https://www.r-project.org/

Utterance Duration

For both participants, periods of voice activity were measured based on the recorded audio. The results are shown in table 10.4 and figure 10.11.

On average, the utterances were rather short, falling between 1 and 3 seconds. The agent's utterances tended to be longer than those of the user, which can be explained by the nature of the scenario. In contrast to the salesperson character who tries to present complex information, the resident character is more likely to utter single words in response. The distributions are shown in figures 10.12 to 10.14.

Speech Alignments

The alignment was annotated as it could be observed from the outside. The behavior model's state was ignored for this part of the analysis because, ultimately, the model is supposed to run in the background and be invisible to end users.

Seven different classes were identified for these alignments, including the distinction between "gap" and "pause" as described in a 2021 review by Skantze [121].

- **Gap:** Silence followed by a change in speaker, or by a repetition of the user's utterance because the system failed to detect it (i.e. the user had released the floor to the agent, but it did not notice).
- **User Solo:** Only the user speaks.
- **User Pause:** Silence between two utterances of the user.
- **User Overlaps:** The user talks over the agent.
- **Agent Solo:** Only the agent speaks.
- **Agent Pause:** Silence between two utterances of the agent.
- **Agent Overlaps:** The agent talks over the user.

As seen in figure 10.15, the salesperson archetype had very little impact on the duration of the different alignment types. Even if the differences were significant, the effect size would be negligible.

However, one unexpected observation was that the agent left very long pauses between individual phrases of its turn. On average, they exceeded the threshold of 3.0 seconds, at which a delay would be considered "long" by the agent. For comparison, Rich et al. [108] set their threshold for valid adjacency pairs to 3.1 seconds of silence. Furthermore, these pauses tended to be longer

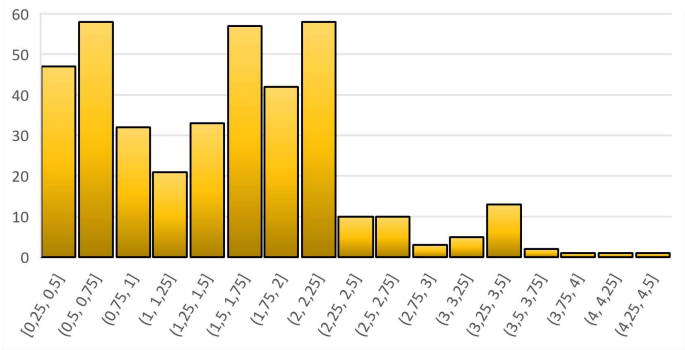


Figure 10.12: Histogram for the duration of the user's utterances.

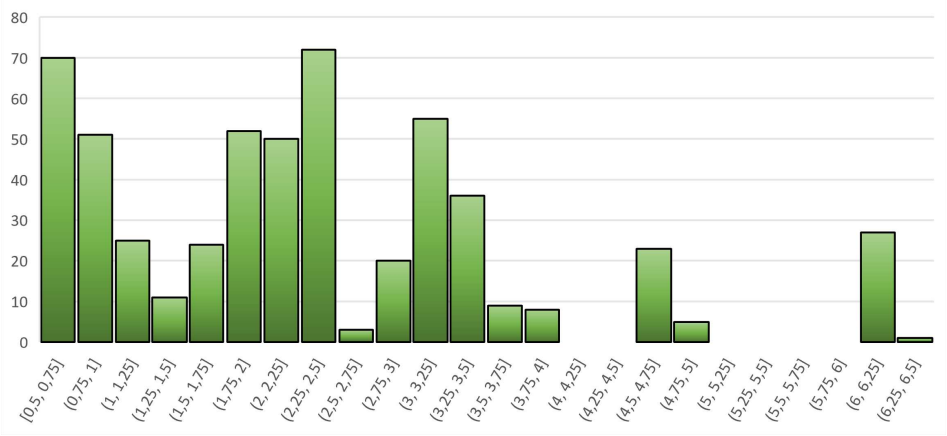


Figure 10.13: Histogram for the duration of the agent's utterances.

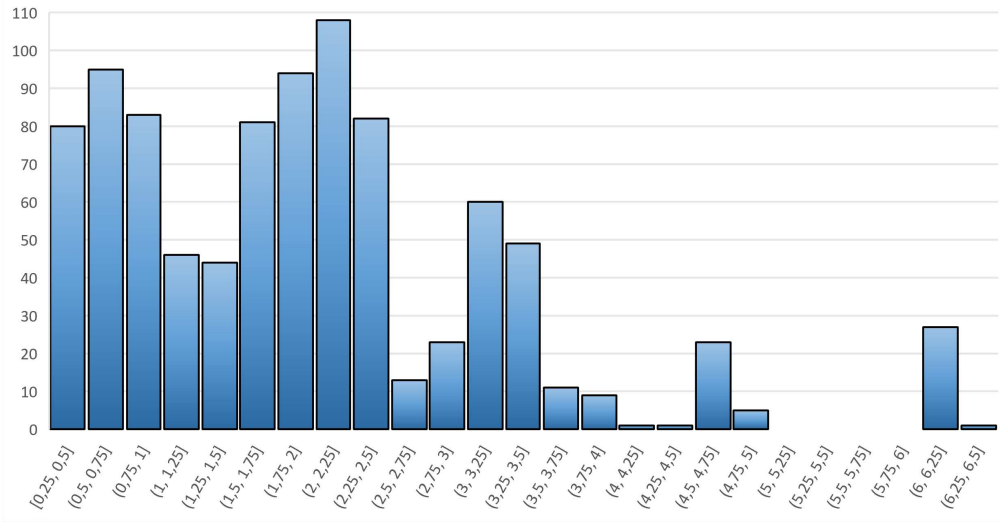


Figure 10.14: Histogram for the duration of both participants' utterances.

than the actual utterances (see previous section). Therefore, they would be very likely to make users impatient and provoke barge-ins if the setup were to be used outside laboratory conditions.

A closer inspection of the prototype revealed that this was a technical issue rooted in the way the state machine was set up. While the issue was eventually resolved, it emphasized the fact that a turn-taking model is only useful when the dialogue manager immediately acts on its decisions.

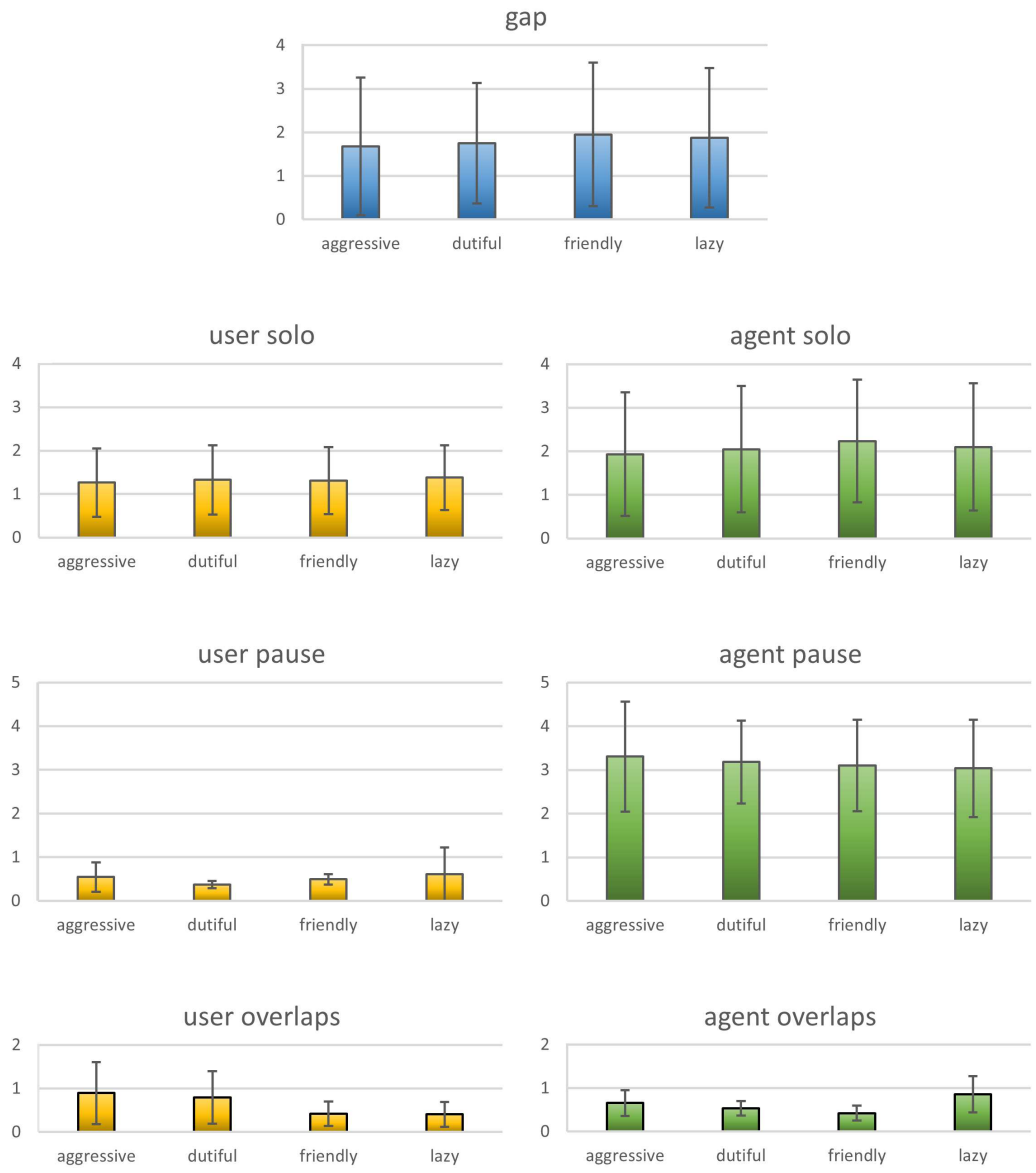


Figure 10.15: Comparison of the alignment durations in seconds that were observed with each archetype.

Conflict Resolution

The semantic content had to be considered to examine the differences in handling turn conflicts. The frequency of certain events was counted for each of the sample dialogues.

- *Seize:* The agent starts talking during the user's turn.
- *Wait:* The agent talks after the user's turn in those cases when other archetypes seize the turn.
- *Hold:* The agent continues talking when the user barges in on its turn.
- *Yield:* The agent ends its sentence early after the user starts talking.

Two ratios were then calculated per sample i, reflecting the agent's response to overlaps depending on who started talking first. Table 10.5 shows the average ratios for each archetype.

$$ratio_{seize,i} = \frac{count(seize, i)}{count(seize, i) + count(wait, i)}$$

$$ratio_{yield,i} = \frac{count(yield, i)}{count(yield, i) + count(hold, i)}$$

agent action	archetype	M	SD
seize vs. wait	aggressive	0.92857	0.18898
	dutiful	1.00000	0.00000
	friendly	0.57143	0.44987
	lazy	0.42857	0.44987
yield vs. hold	aggressive	0.00000	0.00000
	dutiful	0.14048	0.13873
	friendly	0.66905	0.29193
	lazy	0.75000	0.14369

Table 10.5: Observed ratios for an agent archetype choosing the first action over the second one in case of speech overlaps.

Since the ratios were not following a normal distribution, a Kruskal-Wallis rank sum test was performed with the archetype as the independent variable and the respective ratio as the dependent one. Dunn tests were used for

post-hoc pairwise comparison, with the Holm method used for adjusting the p-values.

For $ratio_{seize}$, a significant effect was found with $p = 0.01429$. Pairwise comparisons, however, only showed one significant difference ($p = 0.034$) between the *dutiful* ($M = 1.0, SD = 0.0$) and *lazy* ($M = 0.43, SD = 0.45$) archetypes. The detailed results are presented in figure 10.16 and table 10.6.

Figure 10.16: Relative frequencies of agent actions when it would talk over the user. *Significance:* $* = p < 0.05$

	dutiful	**friendly**	**lazy**
aggressive	0.66055	0.26777	0.09918
dutiful		0.12991	0.03378 *
friendly			1.00000

Table 10.6: Holm-corrected p-values of the pairwise comparison between archetypes for $ratio_{seize}$, the relative frequency of the agent deciding to talk over the user. *Significance:* $* = p < 0.05$.

For $ratio_{yield}$, a significant effect was found with $p = 0.00011$. Pairwise comparisons showed significant differences between *lazy* and *aggressive* ($M = 0.75, SD = 0.14$ versus $M = 0.00, SD = 0.00$: p¡0.001), between *lazy* and *dutiful* ($M = 0.75, SD = 0.14$ versus $M = 0.14, SD = 0.14$: p¡0.05), and between *aggressive* and *friendly* ($M = 0.00, SD = 0.00$ versus $M = 0.67, SD = 0.29$: p¡0.01). The detailed results are presented in figure 10.17 and table 10.7.

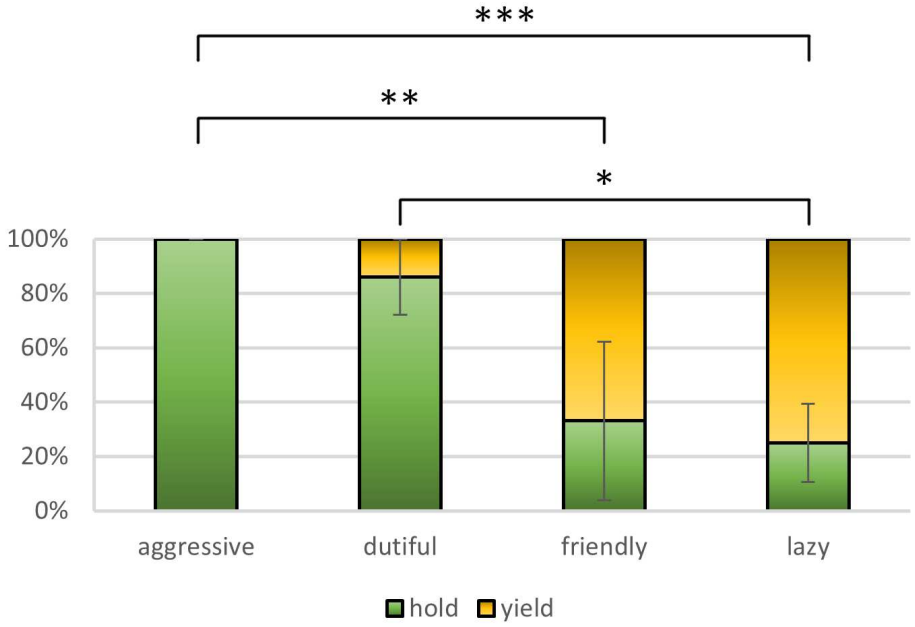


Figure 10.17: Relative frequencies of agent actions when the user starts talking over it. *Significance:* $* = p < 0.05, ** = p < 0.01, *** = p < 0.001$

	dutiful	**friendly**	**lazy**
aggressive	0.57316	0.00284 **	0.00063 ***
dutiful		0.05179	0.01957 *
friendly			0.66507

Table 10.7: Holm-corrected p-values of the pairwise comparison between archetypes for $ratio_{yield}$, the relative frequency of the agent yielding the turn when the user talks over it. *Significance:* $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$.

Gaze Behavior

Eight videos (two per archetype) were inspected more closely regarding the participants' gaze behavior. Unfortunately, the detection of the user's gaze was found to be too inaccurate for a meaningful analysis. Therefore, this section will focus on the agent's gaze.

As intended, the agent was found to avert its gaze when seizing (figure 10.18) or holding (figures 10.19, 10.20 and 10.21) the turn. These patterns were most notable with the *aggressive* and *dutiful* archetypes that were likely to dominate the verbal channel.

When yielding the turn (figure 10.22) or while waiting for its opportunity to

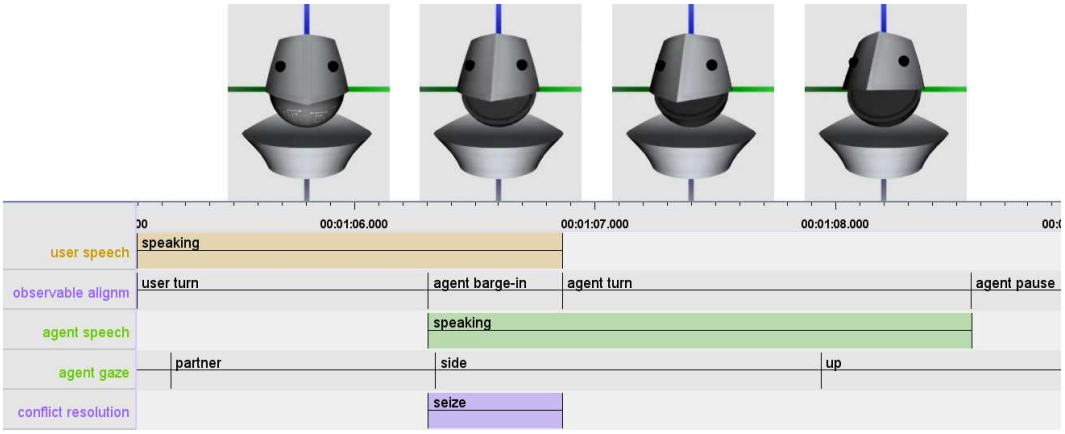


Figure 10.18: Timeline showing example behavior of the "aggressive" archetype while seizing the turn during that of the user.

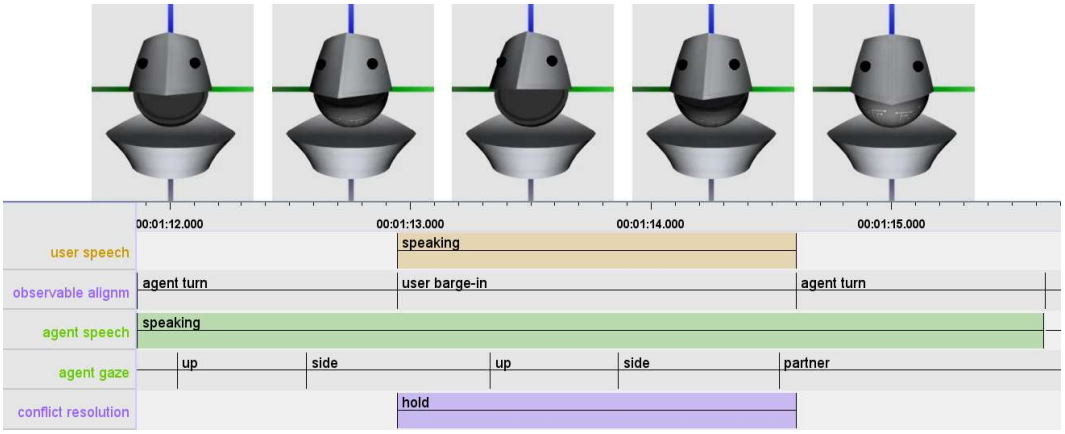


Figure 10.19: Timeline showing example behavior of the "aggressive" archetype while holding the turn during the user's barge-in.

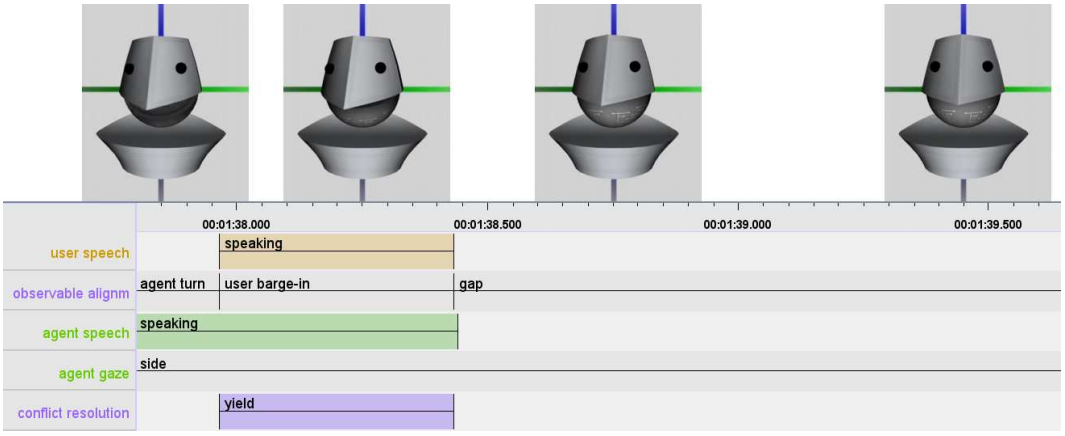


Figure 10.20: Timeline showing example behavior of the "dutiful" archetype holding the turn for a while before yielding to the user's barge-in.

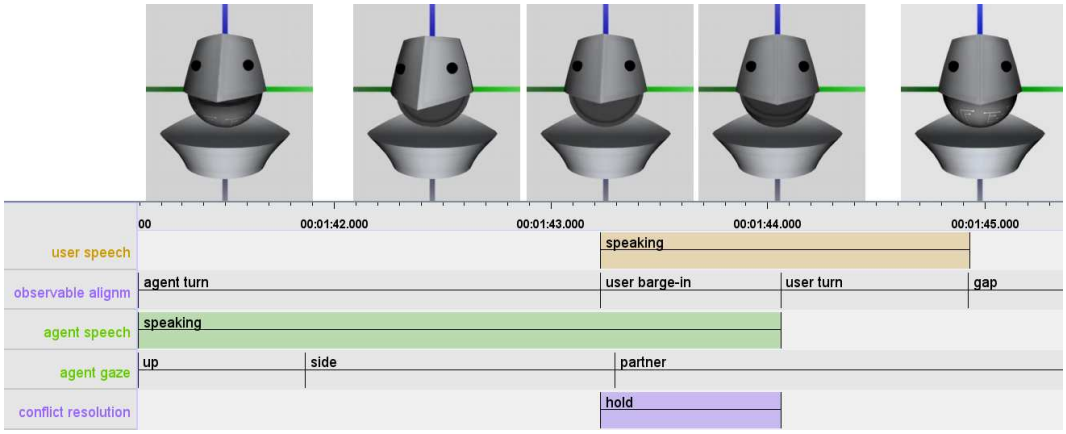


Figure 10.21: Timeline showing example behavior of the "friendly" archetype finishing its turn during the user's barge-in.

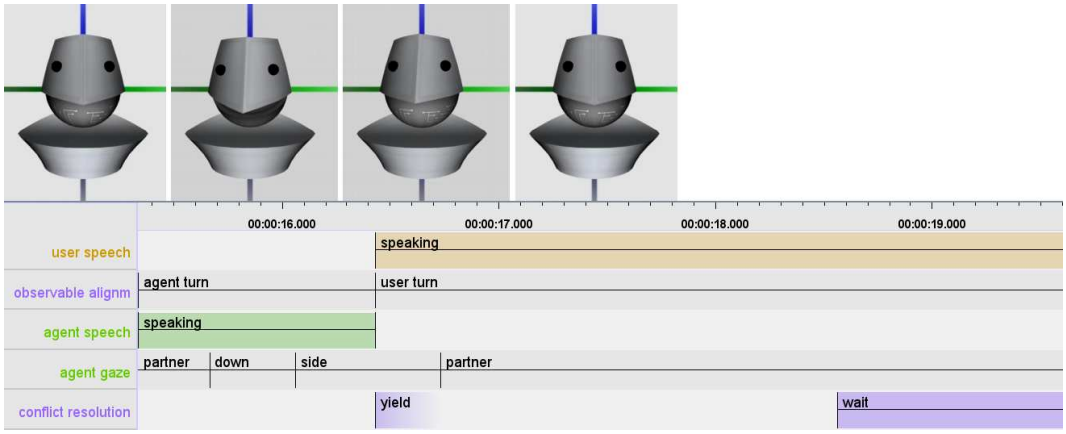


Figure 10.22: Timeline showing example behavior of the "friendly" archetype yielding to the user's barge-in and waiting before the next speaking attempt.

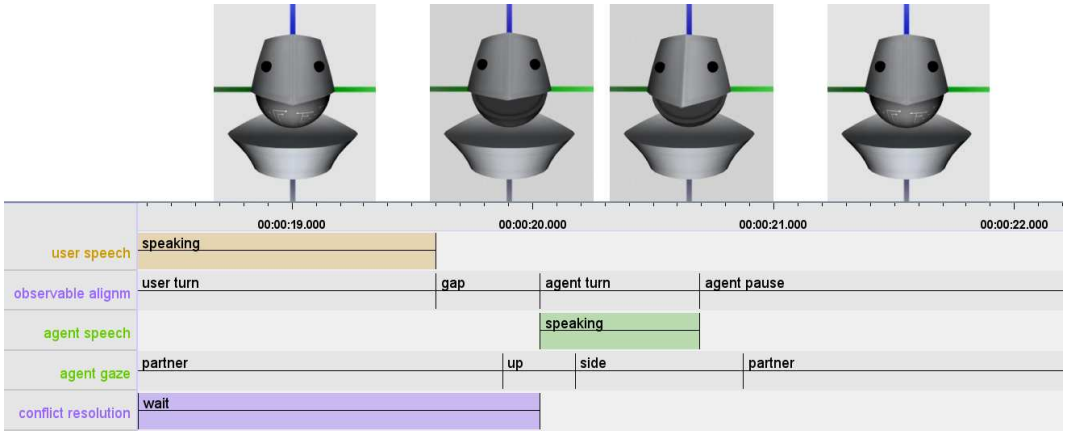


Figure 10.23: Timeline showing example behavior of the "lazy" archetype waiting for the user's turn to end.

speak (figure 10.23), the agent looked at the user in most cases. Occasionally, there was a notable delay between the influence diagram's decision to shift the gaze and the moment when the agent's gaze was visibly changed. This delay can be explained by the smoothing mechanism in the procedural animation. In particular, when the yielding is preceded by a longer phase of averting the gaze, the aversion target is fully activated, and said activation takes longer to decay while the *partner* gaze target needs some time to start outweighing it.

Overall, the inspected samples showed that the agent's gaze behavior depended more on its utterance progress than on its personality.

10.5 Discussion

The turn-taking model succeeded in generating behavior variations that were in line with the literature that it was built on. However, most changes were rather subtle and could only be revealed through a detailed analysis of the recordings.

At the same time, the interactive setup introduced several challenges outside the behavior model. For instance, it was hard to keep the user behavior constant between recording sessions, and the complex interplay of multiple software components would require a similarly complex model to account for delays or inaccuracies.

10.5.1 Behavior Patterns

As it turned out, it is difficult to balance the number of varied personality traits with the need to create distinct agent archetypes. With five traits at five discretization levels each, it is possible to configure 3125 different agent personalities. However, certain trait combinations result in rather similar behavior patterns, especially when they are stripped of their semantic context. More details will follow below.

Objectively, the behavior patterns generated from the configured personalities are in line with psychological literature. However, how human observers will judge them remains to be seen.

Speech Behavior

The emerging speech behavior was very similar for the tested archetypes, and there were barely any differences regarding the alignment of raw voice activity. However, the differences became apparent when looking at the semantic content of the utterances. Taking the actual words and completeness of the agent's utterances into account made it possible to distinguish between cases

in which it insisted on finishing its turn and those in which it yielded to the user interrupting it.

This finding matches comments that study participants made about the non-interactive prototype (see section 9.4.3). In that perception study, several people remarked that they found it hard to determine the characters' personality or interpersonal status without knowing the conversation's topic.

Like humans, the agents controlled by this behavior model can have very different reasons for speaking when they do. For example, the personality traits *Conscientiousness* and *Extraversion* both place a high priority on the goal of speaking one's own contribution to the dialogue. The rationale behind it, however, is different - the highly conscientious agent does so because it wants to fulfill its duty, whereas the extroverted one has an inherent desire to express itself.

Other trait combinations are inherently unsuited for interaction. For example, setting the *Conscientiousness* too low may result in the character not speaking at all, especially when combined with low social interest. While this realistically portrays a lazy, antisocial character, there are few scenarios in which such a personality would be appropriate. Therefore, configuring the salesperson agent to exhibit more extreme behavior would have made it impossible to record a conversation with it.

Gaze Behavior

The agent's gaze depends mostly on its utterance progress. This finding is plausible given how the goal "see" is set up in the influence diagram. Besides the functional relevance, its contribution to the expected utility only depends on one personality-determined factor named "interest". Said factor is calculated as the average of three other factors, each of which is directly derived from a different personality trait or interpersonal attitude component. They are shown in figure 10.24. Since *Affiliation* itself is derived from two personality traits, this means that all personality traits except *Conscientiousness* have the potential of activating the goal "see".

Looking back to the psychological literature, this is in line with the idea that humans may have ritualized the gaze behaviors that are functionally required for conversation. The fact that humans associate differences in gaze behavior with different personality traits might be due to their impact on the amount and timing of speech activity that eventually results in different amounts of gaze.

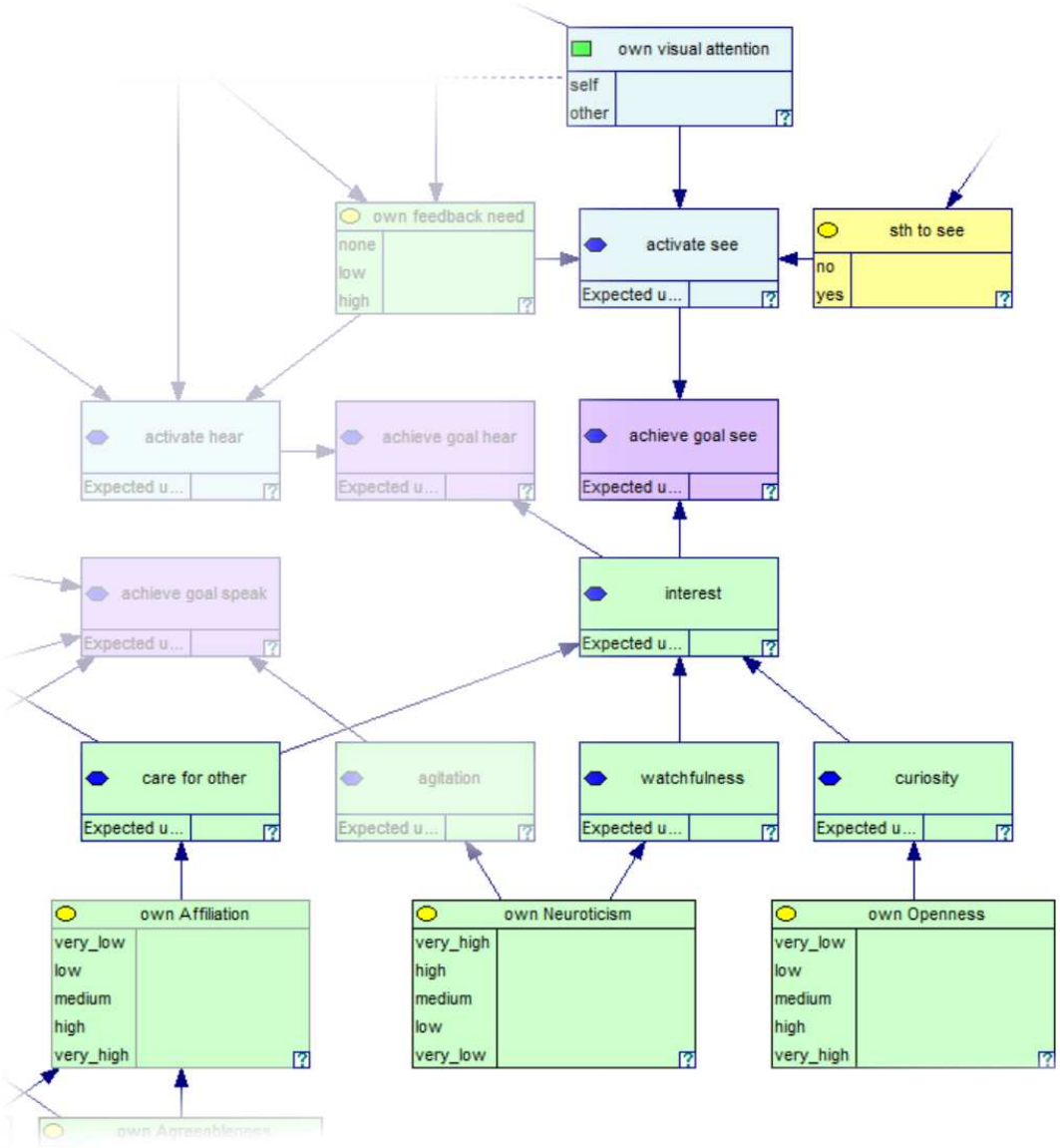


Figure 10.24: Subsection of the influence diagram, showing the personality-derived factors contributing to the activation of goal "see".

10.5.2 Lessions Learned for Evaluating Interactive Setups

Much effort was put into preparing a real-time interactive setup so that the behavior model can be evaluated. However, several challenges emerged in the process, and valuable insights were gained regarding appropriate evaluation methods.

Variations in User Behavior

Experience with recording the samples showed that it is very hard to keep user behavior constant despite several measures being taken to control it. The same person was recorded in every sample, and there was a pre-defined script to follow. Familiarity with the dialogue was very high because the user in question was the same person who authored this thesis, implemented the dialogue setup, and prepared the example scenario. However, reaction times still varied notably. This variation was partially caused by fatigue after numerous recordings and partially by the fact that non-verbal differences are hard to control consciously. Variations in gaze patterns may have influenced the timing on the agent's side, and variations in prosody may explain why speech input and voice activity were not always recognized the same way.

Consequently, a large sample size is required to rule out interference by these behavior variations. While a certain amount of natural variation may be desirable for an interactive demonstration, it can skew experimental results, and great care must be taken when selecting videos as stimuli for perception studies.

Subtlety of Agent Behavior Differences

One major observation was that the behavior variations are very subtle. Turn-taking actions happen on a time scale of a few hundred milliseconds, both for the timing of verbal contributions and the change in gaze direction. They are often only revealed through detailed analysis of the recordings.

This subtlety makes it unlikely that a study participant will pick up on these differences while interacting with the agent in real-time and focusing on their own role in the conversation. In contrast, external observers can pay attention to those details. A perception study based on video stimuli will, therefore, provide more accurate insights into the validity of the behavior model.

Complexity of the Setup

The more complex an interactive setup becomes, the more variables must be considered.

First of all, one needs to ensure that the turn-taking model actually produces the intended surface behaviors. Given that the proposed model is highly complex, this step merits a study of its own. Due to the sheer number of combinations for network observations, it is recommended that this process be automated. For example, a dedicated test application can set the evidence for a specific dialogue context, systematically vary all personality traits (resulting

in $5^5 = 3125$ combinations with the current model), and store the influence diagram's decisions in a format that can be inspected further. Example code for this automated test can be found in B.3.

Furthermore, the experiment's success hinges on the input recognition's accuracy. For example, if the user's voice activity is only detected in 90% of the cases, it becomes hard to see whether the agent's "lack of respect" is due to the configured personality or because it literally did not know that the user was trying to speak.

Therefore, the recommended approach is to evaluate the different components separately and systematically before trying to confront them with naive users.

10.5.3 Technical Bottlenecks

Despite best efforts, several bottlenecks remain within the dialogue setup. They contribute to the delays in the agent's responses and make it difficult to interpret the results. However, this observation also confirms the need for probabilistic reasoning, especially with an approach like that of Bohus and Horvitz [17, 18], who explicitly modeled those inherent latencies.

Input Detection Delay

An earlier batch of test recordings revealed a notable delay between the user's audible speech activity and the visible update of the agent's perception on the associated chance node. A total of 130 changes in user voice activity were annotated on 4 sample recordings.

The average delay between these changes and their reflection in the influence diagram's visualization was 0.357 seconds ($SD = 0.15$). Consequently, there will be discrepancies between the alignments that are observable from the outside and those that the agent believes to produce.

Issues With Debug Visualization

Problems were discovered regarding the display of the system state. Probability distributions and outcome labels in the influence diagram's visualization were rarely updated properly, resulting in impossible states that did not add up to 100% or had two mutually exclusive observations active at the same time. Likewise, the window displaying the received messages from the NLU module was lagging behind. Although it appeared to update in time with the user's speech, the latest result was usually missing. However, that latest result was obviously available to the dialogue manager since the conversation would not be able to proceed without it.

A possible reason could be that large parts of the implementation were multi-threaded to improve performance. It appears that only parts of the influence diagram display were repainted when a change in its state occurred. Forcing Java to repaint specific components in time had been a challenge throughout the thesis, and it could be that a different programming language would be better suited for a real-time application. Alternatively, it could be an option to drop the visualization altogether and simply write the states to a log file that could later be replayed for analysis.

Latencies induced by Visual SceneMaker

Despite code optimization, deactivation of debug visualizations, and running the application on a powerful computer, Visual SceneMaker became notably slower after several dialogue runs. For example, it took longer than usual for switching between the views of different state machine substructures. Visual SceneMaker was restarted several times between recording the samples, but it is likely that this contributed to the observed output delays or compounded the issues with the visualization.

In future work, it would be worthwhile to explore other options for dialogue management. Visual SceneMaker's strength lies in its graphical interface for editing state machines that allows for quick prototyping. However, a more efficient dialogue manager would be needed to handle turn-taking decisions within a fraction of a second.

10.6 Conclusion

Compared to the non-interactive prototype, the turn-taking model changed massively for the interactive version. It was extended to include gaze for disambiguating the participants' intentions, and the representation of the interaction goals was rebuilt from the ground up.

This prototype was then tested in a dyadic conversation between a human following a script and an autonomous agent behaving according to the proposed model. Several versions of this conversation were recorded and annotated to compare the differences in the agent's speech timing and gaze direction.

The primary finding is that semantic context is strictly necessary to notice the differences between personality configurations. The participants' raw speech activity and the resulting alignments did not change significantly between agent archetypes. However, the responses to turn conflicts did, such as stopping in the middle of a phrase when the user barged in as opposed to completing said phrase. This finding should not be surprising, given that the

theory on which the model was built already offered alternative explanations for various surface behavior patterns.

Unfortunately, with the increased complexity of the setup, there was also an increase in delays that were not caused by the behavior model itself. Latencies and inaccuracies were observed in the input pipeline, similar to the issues described by Bohus and Horvitz [17]. The work done in this thesis confirms that the interplay of so many different software components requires an equally complex model to account for the numerous sources of noise and errors.

Part IV

Outlook

Chapter 11

Contributions

11.1 Introduction

Personality plays an important role in creating believable conversational agents, even more so if their design invites humans to anthropomorphize them. To create a consistent character, its verbal and nonverbal behaviors must be in line with the personality that it is supposed to convey. Consequently, this thesis focused on how the turn-taking behavior of an ECA can be varied to express different traits, using a decision-theoretic approach to model idealized human reasoning.

Behavior generation based on statistical models often suffers from the problem that its decisions are intransparent and not always in line with human reasoning. For example, Lapuschkin et al. [74] demonstrated how image classification can base its decision on irrelevant and even misleading clues such as a photographer's watermark or uniformly colored padding at the edge of a photo. In contrast, a decision-theoretic model relies on pre-existing knowledge, such as established findings from psychology, to structure the decision process while also incorporating statistical data in the form of conditional probabilities. Therefore, this approach was chosen to provide an alternative to current machine learning technologies.

This chapter will sum up the contributions to the scientific field. They are sorted into methodological and technical contributions.

11.2 Methodological Contributions

Psychological findings were thoroughly reviewed to build a behavior model that was properly grounded in existing theories. The connections made here

provide a solid foundation for generating turn-taking behaviors in this thesis and in future works that will build on it.

11.2.1 Connection between Personality-related Models

Established models for personality, interpersonal attitude, politeness, and emotions were studied. Intersections between the underlying concepts were identified and backed up by literature. These intersections were then used to unify the findings about social interactions and communicative signals.

Being able to convert one model to another, such as the Big Five personality traits to the Interpersonal Circumplex [39], ensures that researchers in the human-agent community can make the most of the existing findings from psychology. For example, communicative behaviors that were examined with regard to Extraversion can be reframed in the context of Brown and Levinson's Politeness Theory [20] via the definition of interpersonal dominance. Consequently, it becomes easier to trace behaviors back to a limited set of factors from which an agent's behavior can plausibly be derived.

11.2.2 Relating Personality Models to Communicative Goals

Existing goal taxonomies [29, 129], politeness theory [20] and the OCC2 model [98] were examined in order to identify concrete goals that would be reflected in an interlocutor's behavior. Two core ideas emerged from that. First, different cultures judge communicative acts differently when if comes to face threats [20, 126]. Second, not all goals are in focus at the same time.

The conclusions from these core ideas were that the personality traits influence the degree to which an artificial character "cares" about achieving the available goals, while the functional state of the conversation determines whether a goal even needs to be considered at a given moment.

Each of the personality-based weights was associated with a quality that is commonly used to define the trait in question, such as a sense of duty for conscientiousness or curiosity for openness.

11.2.3 Relating Communicative Goals To Behavior

The desire for information was identified as the factor that could explain most behavior patterns in the context of turn-taking. Every communicative goal was eventually reduced to a wish for obtaining additional information, avoiding it, or helping the interaction partner fulfill their own information need.

Consequently, the turn-taking model was built on the idea of selecting an attention target for each modality. Those targets then determine whether

a communicative channel is opened or closed, limiting the actions that an agent can perform at any given moment and thus regulating the amount of information being passed between them.

11.2.4 Decision-theoretic Turn-taking Model

Based on the findings in the reviewed literature, a decision-theoretic approach was proposed for reasoning about the agent's turn-taking behavior.

The agent's personality traits, as well as its beliefs about the conversational context and the interlocutor's actions, were mapped to chance nodes in a Bayesian network. Factors involved in prioritizing the goals, such as specific facets of the Big Five traits or the relevance of a goal at a given moment in the conversation, were represented by basic utility nodes. The values specified at those nodes were then combined into multi-attribute utilities, reflecting how useful a specific attention target would be for achieving the associated goal, as well as how much the configured agent personality was interested in achieving it in the first place.

The agent's behavior was then derived from decisions about its attention target, updated whenever new information about the participants and dialogue context became available. After calculating the expected utility for each target and modality, the agent's verbal and visual attention was finally set to the targets that best fulfilled its goals.

11.3 Technical Contributions

Besides developing the turn-taking model itself, a software ecosystem had to be implemented to connect it to a dialogue application and facilitate prototyping with different agents.

11.3.1 Participant Framework

On the side of the dialogue manager, participants were implemented in a modular, extensible way. Computer-controlled agents are not allowed to share knowledge directly in order to simulate the interaction with a human. Instead, they are forced to rely on the same messages that they could also receive from the sensors detecting user input.

Communication between all participants, both humans and ECAs, is managed via a central message hub. Messages carry communicative acts that are inspired by the DiAML standard [60, 101, 22] and either inform the interlocutor of the sender's nonverbal behavior or advance the dialogue with verbal contributions.

Context information, such as the gaze direction of the interlocutor or the delay since the agent's last speaking attempt, is tracked separately from the influence diagram. It is only discretized when a decision is needed, specifically whenever a new message is received or when the agent moves on to the next phase of delivering its utterance.

Finally, the decision of the influence diagram determines when the current speech command is forwarded to the agent's behavior realizer. It is stalled until the agent's verbal attention shifts towards itself, and if the attention should shift towards the interlocutor before completion, the speech command is canceled. This way, delays or overlaps emerge in real-time without the need to plan ahead.

11.3.2 RobotEngine Framework

The RobotEngine framework was developed as a uniform interface between different control applications, such as Visual SceneMaker or a simplified Wizard-of-Oz experiment control panel, and different artificial agents, such as the Klappmaul character, the RoboKind R-50 Zeno, or the Robopec Reeti. A standardized messaging protocol decouples the behavior realization from its semantic meaning, making it easier to reuse and reconfigure existing setups.

All high-level scheduling, such as the timing of speech commands or the selection of gaze targets, is done by the control application. Unimodal commands are sent to an agent-specific RobotEngine implementation where they are mapped to the necessary API calls. The RobotEngine handles low-level conflicts, such as simultaneous movement commands for the same servo motor, and monitors the execution progress. Said progress is translated to one or more status messages that are sent back to the control application.

Additionally, several requirements for smooth turn-taking were identified while connecting the different agents to this framework.

- **Asynchronous Behaviors:** The agent must support the parallel execution of speech and animation commands. Otherwise, it would be impossible to send the gaze signals used for coordinating speaking turns.

- **Progress Monitoring:** The agent must provide information about the execution progress, such as bookmark events from the TTS service or a notification when an animation has finished. Otherwise, the agent will not know when it succeeded in speaking the MNI or where it is looking currently.

- **Canceling Commands:** The agent's API must expose stopping commands for started behaviors. At the very least, it must be possible to

cancel speech output that is already in progress. Otherwise, the agent cannot yield the turn when the turn-taking model demands it.

The RobotEngine framework is available on GitHub at https://github.com/kjanowski/RobotEngine. The repository contains the core classes for Java, C#, Python 2, and Python 3.

Since most robots and graphical agents rely on proprietary software libraries, their RobotEngine implementations cannot be distributed publicly. However, the main repository contains classes for controlling a Unity character in combination with the third-party asset "RT-Voice" by crosstales[1]. A separate repository exists for the Java-based "Klappmaul" agent, a reference implementation used for testing both the RobotEngine framework and the turn-taking model. This agent can be found at https://github.com/kjanowski/Klappmaul.

11.3.3 Proof of Concept

Two example applications were implemented to showcase the real-time generation of turn-taking behavior based on the decisions of the influence diagram.

The first one showed a conversation between two computer-controlled characters, with the human merely observing their behavior. The characters in question were realized as separate processes. Communication between them was limited to what they could plausibly know about a human interlocutor. A perception study was conducted to confirm that the varied personality trait, *Extraversion*, resulted in different speech timings while influencing the perceived *Agreeableness* and interpersonal *Status* as described by psychological literature [80, 79, 39]. The results were published at the International Conference on Autonomous Agents and MultiAgent Systems in 2019 [64].

In a later phase of the thesis, an interactive human-agent conversation was implemented with state-of-the-art input recognition. The Retico framework[2] [88] and Rasa[3] are used for incremental speech parsing while MediaPipe[4] provides the user's current gaze direction.

Sample interactions were recorded and analyzed to see if the turn-taking model generates suitably distinct behavior patterns and whether it runs efficiently in a real-time setup. The results showed that the semantics attached to the timing decisions played a major role in telling the personality configurations apart. The number and duration of overlaps, silence, or single speaker activity were not sufficient to distinguish between an *aggressive*, *dutiful*, *friendly*,

[1] https://assetstore.unity.com/packages/tools/audio/rt-voice-pro-41068
[2] https://github.com/retico-team
[3] https://www.rasa.com/
[4] https://google.github.io/mediapipe/

and *lazy* character archetype. However, those archetypes did differ in terms of conflict handling, as revealed by taking the utterance completion into account and directly comparing the agents' timing for specific sentences.

11.4 Conclusion

This thesis made both methodological and technological contributions to the scientific field.

Psychological literature was systematically reviewed and sorted to develop a decision-theoretic model for the turn-taking behavior of an ECA. The connections between personality, interpersonal attitude, and politeness theory were summarized, and attention was identified as the foundation of most turn-taking behaviors. Based on the definition of personality traits and interpersonal attitudes, factors were selected for prioritizing the agent's goals in a turn-taking context.

Besides developing the proposed turn-taking model as an influence diagram, the required software was implemented to test it in actual dialogue applications. This software environment encompasses two frameworks. One is attached to a dialogue manager, handles the semantic communication between an arbitrary number of computer-controlled or human participants, and regulates the agents' behavior based on the influence diagram. The other serves as a uniform interface to different graphically embodied agents and social robots, separating their implementation details from the interaction logic and thus ensuring that the turn-taking model can be used with any agent connected to this framework.

Two applications were set up to evaluate the presented approach. The first one had two separate autonomous agents talking to each other as if each one was talking to a human user. Specifically, they were forced to infer the interlocutor's intended role (speaker or listener) from only the observable voice activity. Although the model was very simplified, taking only the most salient personality traits into account, a perception study [64] confirmed that it succeeded in generating the intended behavior variations.

For the interactive prototype, the psychological literature was revisited, and the turn-taking model was rebuilt more rigorously to cover all five traits of the Five Factor Personality Model. The goals that the agent sought to achieve through its behavior were made less abstract and linked to the need for obtaining or providing information. The preliminary evaluation revealed that, as more personality traits offer alternative reasons for outwardly similar behavior patterns, the semantic content of the conversation becomes crucial for telling the personality configurations apart. The generated behaviors are in

line with the theory on which the turn-taking model was built. Nevertheless, more studies will be necessary to see how human observers will perceive these often very subtle differences.

Chapter 12

Future Work

12.1 Introduction

The nature of science is that every answer gives rise to new questions. Often, one can only start asking the right questions after becoming familiar enough with a given subject. Implementations can always be improved, and every failed test run can teach new lessons for the next iteration. However, for every project, there comes a point to draw the line and wrap it up. This thesis is no different.

Therefore, this chapter will look back on the issues that are still unresolved and point out potential directions for future work. The following section will summarize several limitations of the turn-taking model itself, the technology to which it is connected, and the scenarios in which it has been tested so far. After that, there will be a section on possible improvements of the technology, as well as its application for further research on human communication or the development of adaptive human-agent interfaces. The final section will conclude not only this chapter but the thesis as a whole.

12.2 Limitations

Turn-taking is a far more complex subject than what can be covered in a single thesis, especially when it comes to multimodal communication with real humans. Consequently, not all aspects mentioned in the literature could be included in the presented behavior model. The proof-of-concept implementation also brought some challenges to light, both on the technical side and that of the application scenario.

12.2.1 Model Limitations

The current version of the turn-taking model only scratches the surface of what could be represented. In particular, it does not use the underlying Bayesian network to its full potential, and only a subset of the known turn-taking signals are included so far.

Modalities

When reviewing the literature on turn-management cues, Skantze [121] listed a wide range of signals in several different modalities. This thesis only covered a fraction of those in its turn-taking model - specifically, voice activity and gaze direction.

For example, raising the volume is a known strategy for defending one's speaking role against the person trying to take over, whereas lowering it indicates a willingness to yield [69]. The agent's speech volume could be added as another action on which the influence diagram needs to decide, while that observed from the user could provide additional evidence for inferring their intention. The same goes for the pitch patterns associated with the end of a turn [69].

Overall, the underlying principles of Bayesian networks will help keep the complexity manageable as the model becomes more detailed.

Uncertainties

In its present version, the turn-taking model only considers the uncertainties that are directly linked to the user's intentions and information needs. However, many more uncertainties are involved in dialogue applications, especially when it comes to real-time user interaction.

For example, Bohus and Horvitz [17] explicitly modeled several latencies that were to be expected within their system's input and output pipelines. In the context of this thesis, attempts were made at predicting future alignments (see figure 12.1). However, they turned out to require a strong enough hypothesis regarding the user's intention and the length of their verbal contribution.

Another aspect that should be incorporated in future versions of the model is the precision of the input recognition. Bayesian networks are ideal for representing the true world state given a particular sensor input, similar to the reliability of mechanical or medical tests that Neapolitan models in several examples [94]. Therefore, better results could be obtained by considering the accuracy of the available voice activity detector or the gaze direction classifier.

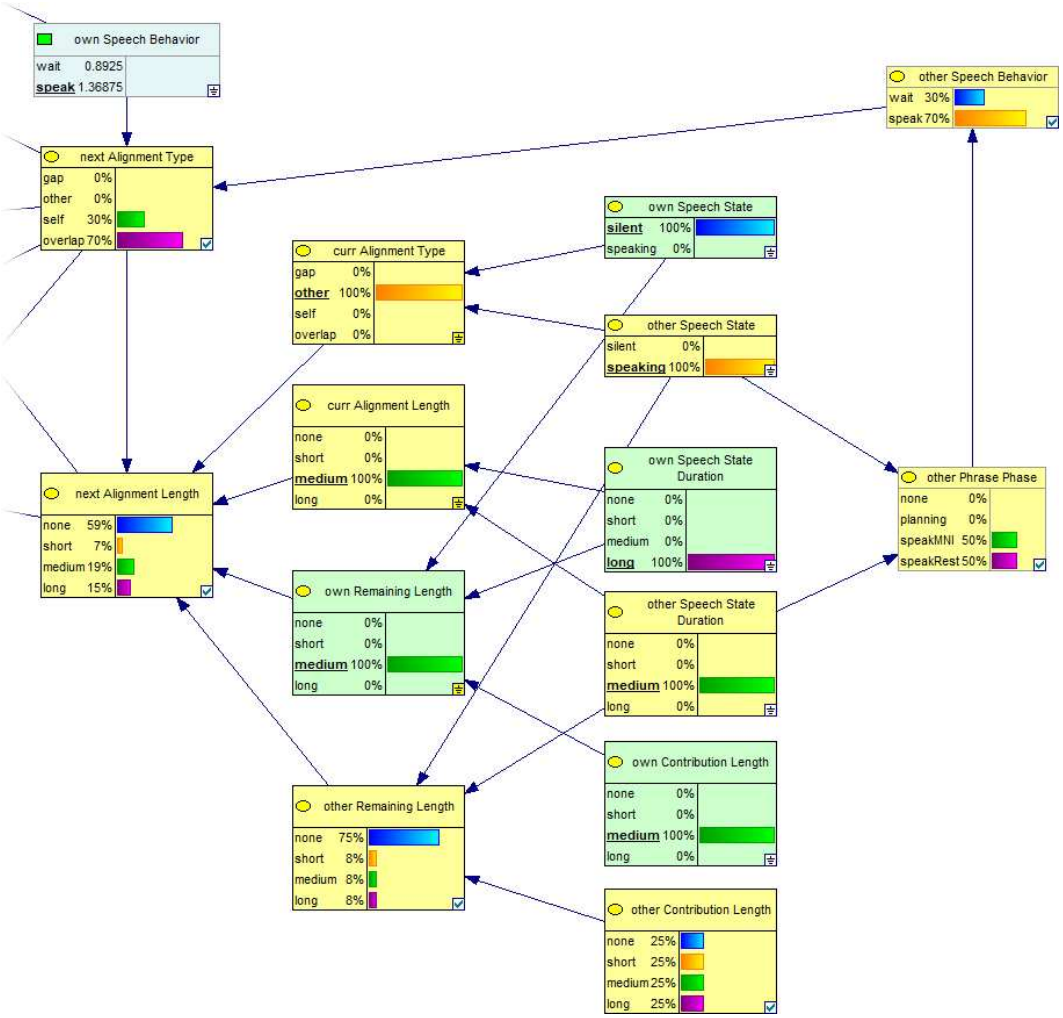


Figure 12.1: An excerpt of a discarded prototype, showing an attempt at predicting the alignments resulting from the participants' turn-taking decisions.

12.2.2 Technical Limitations

Real-time interactive setups depend on efficient computation. While the proposed model avoids complex predictions about future events, it still involves many calculations that are triggered with a high frequency. In its present form, the turn-taking model comprises all of the major personality traits and several fundamental, domain-independent goals. It performed well on a state-of-the-art laptop, but it should be noted that it only covers a subset of the related modalities. For example, the volume of the voice or turn-requesting gestures have not been considered so far, and extending the model with them would add several more steps to the calculation of the expected utilities.

Latencies

As the interaction setup becomes more complex, there is also an increasing chance to introduce additional latencies that are unrelated to the turn-taking model's decisions. Detailed parallel state machines, real-time input processing, and speech synthesis were identified as potential sources of such delays. These observations are in line with related work. For example, Bohus and Horvitz [17, 18] pointed out similar delays in their virtual quizmaster application and explicitly included them in their reasoning about the agent's behavior timing.

However, there is the potential of extending the influence diagram to represent such latencies as well. As Bohus and Horvitz showed, they can be represented with a probability distribution for observing delays with specific durations. Therefore, more chance nodes could be added to the influence diagram to more accurately reflect the consequences of the agent's speech or gaze timing.

Hardware Requirements

Most of the latencies could be alleviated by upgrading to a recent, more powerful computer system. Table 12.1 compares the specifications of the Samsung notebook used for the non-interactive prototype in 2019 to those of the mobile Lenovo workstation used for the interactive version in 2023.

	Samsung Notebook Serie 5 Ultra	**Lenovo ThinkPad P14s 2nd Generation**
CPU	Intel Core i7-3517U 2 cores, 1.90 GHz	AMD Ryzen 7 PRO 5850U 8 cores, 1.90 GHz
RAM	8 GB DDR3	48 GB DDR4
GPU	NVIDIA GeForce GT 620M 1 GB DDR3	AMD Radeon Pro Graphics 4 GB DDR4
WiFi	802.11 abg/n, max. 300 Mbps	802.11 ax, max. 2.4 Gbps
Operating System	Windows 8 (64 bit)	Windows 10 (64 bit)

Table 12.1: Hardware specifications of the laptops used at the beginning respectively at the end of this thesis.

Unfortunately, the hardware requirements make it unlikely that the turn-taking model can be run directly on current robot platforms. Most of these have limited processing resources due to additional requirements regarding their overall size, weight, or form factors. For mobile robots, said requirements

also affect the battery capacity, which in turn limits the power supply to its processing units. As for virtual agents that need to run on the same machine, the graphical display can take away from the available resources, so a dedicated GPU is a vital prerequisite.

To some degree, it is possible to offload input processing and agent control onto different machines. For example, the social robots used during this thesis only receive commands for the desired surface behaviors and take care of scheduling their own low-level resources. The dialogue manager and the turn-taking model are running on a separate machine that is also responsible for processing the user input.

12.2.3 Scenario Limitations

Expressing personality via turn-taking variations is only possible when the interaction topic provides opportunities for acting differently. There must be a sufficient probability of longer sentences that invite interruptions or overlaps as opposed to straightforward question-and-answer exchanges. Consequently, a typical home assistant scenario with commands and suggestions will have little to gain from this turn-taking model, and a simple detection of user barge-in may be sufficient for an agent that provides long explanations.

However, the personality-specific handling of turn-taking conflicts can add another layer of realism to training simulations. In these settings, the agent is not expected to cooperate with the user, and therefore, turn conflicts such as interruptions or awkward silences are intentional parts of the interaction design. In these cases, the simulated personality serves as a way to adjust the training's difficulty and provide the trainee with a wide range of example situations.

How well the turn-taking model performs in a more realistic scenario, such as a training simulation or a character-driven game, remains to be seen.

12.2.4 Evaluation Limitations

So far, the interactive prototype has only been tested with one single user, specifically the author of this thesis. The analysis of the recorded sessions already provided several valuable insights, but these tests are only the first step toward a proper evaluation.

An interactive system poses major challenges when it comes to reproducible behavior and controlling for interfering variables. For example, the sensor accuracy may vary with the time of day, the color of the participants' eyes, or the base pitch of their voice. Experience has shown that keeping the user's behavior constant is very hard, even when they follow a fixed script and are highly

familiar with the scenario. Consequently, a large number of study participants will be necessary to obtain meaningful results.

Furthermore, the preliminary evaluation focused on objective measures, such as the agent's reaction to speech overlaps or observable gaze sequences. To validate whether the personality is expressed appropriately, the agent's behavior needs to be judged by humans. This subjective evaluation could be done by either the interacting person or an external observer. While the former option is closer to the behavior model's intended use, the latter has the advantage that the observer can easily focus on the interlocutors' behavior.

12.3 New Directions

Several questions emerged during this thesis that could not be answered within its scope. Some topics, such as tailoring an agent's personality to the user's requirements, were explored to a certain degree. However, many were eventually dropped because they would have opened up too many side projects.

The turn-taking approach presented here will also be an important step toward developing agents with more human-like conversational skills. Such agents will be useful for various research purposes in both human communication and human-computer interaction.

This section summarizes the most salient directions for future research and improvements to the implementation.

12.3.1 Theory

Over the course of this thesis, several vague or conflicting theories were found regarding communicative behavior in humans. For example, no definitive answer was found on what determines the time that a person spends looking at another, which personality is linked to a particular set of goals, or how much overlapping speech is tolerable in which context.

The psychological literature was found to provide mostly general tendencies, whereas concrete numbers were mostly found in computer science works. An ECA that displays behavior in line with those theories and numbers will help greatly with filling in the gaps, allowing for gradual refinement of the theories and the systematic search for the related durations, frequencies, or ratios.

Validating the Personality Expression

The first step, of course, would be to conduct a more in-depth experiment to evaluate the interactive prototype. After addressing the limitations of the cur-

rent implementation and the chosen dialogue scenario, the agent's turn-taking behavior should be presented to an appropriate sample of people who will judge its personality and interpersonal attitude using a validated questionnaire.

Since real-time interaction with humans introduces several interfering variables (see section 10.5.2), it is recommended to run a video-based perception study first. If the personality is indeed perceived as intended, a follow-up study can be planned to confront users directly with an autonomous agent. At the moment, the interactive prototype is being revised and adapted for use with a Robopec Reeti, based on the lessons learned during this thesis.

Studying Personality Perception

Unlike humans roleplaying a particular personality, a computer-controlled character is guaranteed to perform consistently across sessions. This consistency opens up new possibilities for studying the factors that influence human observers' judgment.

For example, gender stereotypes may color the degree of dominance or affiliation that is considered acceptable for an agent. Based on its role in the scenario, an agent might appear confident, arrogant, or impudent. A red robot might be perceived as more aggressive or emotional than a blue one despite showing the same behavior. Depending on the topic, humans could attribute different motivations to a character when it exhibits a particular behavior.

The turn-taking model described here contributes an important piece to this puzzle. The underlying personality can be configured to present humans with a wide range of interaction partners that exhibit human-like turn-taking capabilities. Combined with a naturalistic virtual human or android, this can also be expected to increase immersion and yield more realistic results.

12.3.2 Technology

During this thesis, several pain points became apparent in currently available agent architecture, along with ideas for addressing them. At the same time, certain technologies - most notably, generative conversational AI - only emerged during that time and have not yet been combined with the proposed approach for modeling turn-taking behavior.

Dialogue Manager

The results obtained from using this behavior model are only as good as its connection to the software that manages the surrounding interaction.

For the interactive prototype (see chapter 10), a considerable part of the scheduling logic was implemented using hierarchical and parallel finite state

machines. While this made it easy to test certain approaches quickly, it also had several downsides. For example, many scene graph patterns, such as those for retrying interrupted sentences, had to be copied manually, increasing the risk of mistakes in the interaction flow. The synchronization between the state chart selecting the next utterance and the one passing it to the agent was rather complicated, introducing a bug that caused unintentional delays and was only discovered very late.

For future versions, those parts of the scheduling logic should be translated into regular code. Besides avoiding errors and hiding those technical details from the interaction designer, this conversion is also expected to increase computational efficiency.

Furthermore, after cleanly separating the scheduling logic from the interaction flow, it will become easier to explore options for embedding the turn-taking model in different dialogue management frameworks. For instance, the rise of conversational AI solutions begs the question of how those could benefit from more human-like turn-taking behaviors.

Generative Conversational AI

Recently, generative conversational AI such as ChatGPT[1] has spread through all kinds of applications. The flexibility of LLMs and the quality of the generated text make this approach attractive for social agents that need to have a consistent conversation with a human.

It would be worth exploring a connection between ChatGPT and the presented turn-taking model. Such a setup would be rather straightforward. It would require the following:

- **Speech recognition:** The raw speech input can be detected using the wav2vec [13] module that is included in the Retico framework[2] [88].
- **NLU component:** It is likely that ChatGPT will be able to process partial input on its own, so the transcribed words could be passed to it directly.
- **Response processing:**
  - **Utterance buffer:** Whatever ChatGPT would answer to the (partial) input must be stored until the turn-taking model allows the agent to speak.

[1] https://openai.com/chatgpt
[2] https://github.com/retico-team

- **Agent participant:** An implementation of the agent participant (see section 7.5) that tries to speak the currently buffered sentence(s) when the influence diagram allows it.
- **MNI detection:** A way to determine when the agent has said enough to be understood and can move on to the next sentence. A second instance of ChatGPT could possibly do this.

Agent Platforms

Over the course of this thesis, several requirements have been identified that ECAs must fulfill before they can display the generated turn-taking behaviors.

As explained in section 8.2, notable prerequisites are the parallel execution of actions in different modalities, the ability to cancel those actions after they were started, and sufficiently detailed feedback about the execution progress. Knowing these requirements will help implementing agents in a way that they can actually benefit from a sophisticated turn-taking model.

12.3.3 Application Scenarios

This thesis provides the foundation for more complex use cases that call for a consistent and configurable personality model. The turn-taking approach presented here can be combined with other types of behavior generation, such as mapping simulated emotions to facial expressions [5] or changing the linguistic style to match the agent's personality [110].

For example, it could be applied in the context of training simulations. Different archetypes for virtual roleplay partners can be created easily by changing the underlying personality traits, and deriving the behavior from those traits ensures that the agent acts in line with the intended characterization. Previous research by Gebhard et al. [45] showed that different agent personalities contribute to the challenge of roleplay situations, as showcased with a job interview scenario. It is easy to imagine how such variable challenge levels can augment simulations for other domains, such as negotiation [38, 135, 137] or medical training [96].

Another use case would be teaching motivational interviewing strategies [92] by simulating clients with varying levels of compliance (agreeableness) and discipline (conscientiousness). Such a scenario was considered for the interactive prototype (see chapter 10) to use synergies between this thesis and the research at the chair of Human-centered Artificial Intelligence. Unfortunately, it was hard to identify opportunities for interruptions or undesirable silence in this context, so it was less attractive for showcasing turn-taking variations. Implementing such a conversation would also have required considerable amounts

of domain knowledge regarding the habit change at the center of the conversation. It should be noted that Yang et al. [140] did examine interruptions in the context of the PANORAMA project. However, they have not yet applied their approach to the motivational interviewing domain, possibly for similar reasons. It would be interesting to apply both approaches in a related scenario and compare the results obtained with their statistical timing prediction to those obtained with the decision-theoretic model presented here.

12.3.4 Adaptive Personality

The Participant Framework (see chapter 7) was built with extensible agents in mind. One such extension could be to equip an agent participant with machine learning in addition to the influence diagram. For example, reinforcement learning can be enabled by adding a Q-table to an *InterruptibleAgentParticipant* and implementing methods for reward calculation and action execution. The situation parameters that the basic *AgentParticipant* tracks can easily be mapped to a state label, with additional options in the parameter configuration specifying which ones need to be included.

This way, the agent's turn-taking behavior could be tailored to a particular user's preferences or requirements. Changing the underlying personality configuration instead of the behavior itself would ensure that the result remains consistent.

The separation between the situation parameters and the decision-theoretic model makes it possible to alter the influence diagram's parameters based on the learned policy. The state space used for learning and the observations in the influence diagram need not be the same, which saves computation time and reduces the number of examples that are needed in the training data. It also keeps the turn-taking model independent from the interaction domain.

For example, the agent could exhibit more extroverted behavior if it discovers that a particular user responds better to messages delivered that way. However, the response, such as adherence to a diet plan or change in mood, would not be part of the turn-taking model itself. Instead, the learning algorithm would tell the turn-taking model to assume different personality traits for the agent and make it behave accordingly. After some time, the learning algorithm would evaluate if that change brought it closer to the long-term goal, such as the agent being perceived as competent or entertaining.

12.4 Conclusion

Nowadays, we are surrounded by an increasing number of conversational agents that not only require a certain social competence but also personalities to match their roles. Such a personality can be expressed in many ways, including emotional responses, linguistic style, or tone of voice [66]. This thesis focused on expressing it through the agent's turn-taking behavior, inspired by the personality traits and interpersonal attitude that humans tend to associate with specific speech patterns.

A decision-theoretic approach was chosen to model idealized human-like reasoning, striking a balance between the "gut feeling" on which people tend to rely for quick reactions and the cold logic that they tend to expect from machines. Furthermore, the graphical representation in the form of an influence diagram was an attractive alternative to the intransparent statistical models that are popular in state-of-the-art dialogue systems.

The turn-taking model was embedded in two different dialogue setups. One tested the core ideas by letting two autonomous ECAs talk to each other, acting only on the information they would have obtained from a human speaking into a microphone. The second dialogue setup provided real-time interaction between a human and an ECA that was controlled by a more comprehensive model. The behavior variations observed with the latter prototype were rather subtle and mostly found through a detailed analysis of sample recordings. While they were in line with the expected patterns, it remains to be seen if they are sufficiently different for human observers.

At the time of writing, there are already plans for improving the interactive application and applying the lessons learned during the thesis to the design of several follow-up experiments. The next step will be to determine several personality configurations with distinct behavior patterns that are to be presented in a video-based perception study. If participants' perception of these archetypes aligns with the configured traits, another study can be performed with the interactive setup.

Besides conducting more thorough evaluations and exploring different application scenarios, there is also much potential for expanding the presented model. Turn-taking is a complex topic, so the behaviors covered in this thesis are only the tip of the iceberg. So far, only a fraction of the involved nonverbal signals were incorporated, and tests with the interactive setup confirmed that it needs to take more uncertainties into account than just those concerning the user's cognitive state.

In the end, every interface developer will need to trade off the realism of a behavior model against its simplicity. A comprehensive, personality-based turn-taking model may not bring a notable benefit for everyday scenarios such

as home assistants or receptionists. In contrast, immersive training simulations or interactive storytelling can benefit greatly from consistent behavior generation based on psychological theories.

It will be exciting to see what the future holds for computer-controlled characters and how much said future has come closer through this thesis. Perhaps the artificial life humans have wanted to create since the dawn of history is within our reach sooner than we thought.

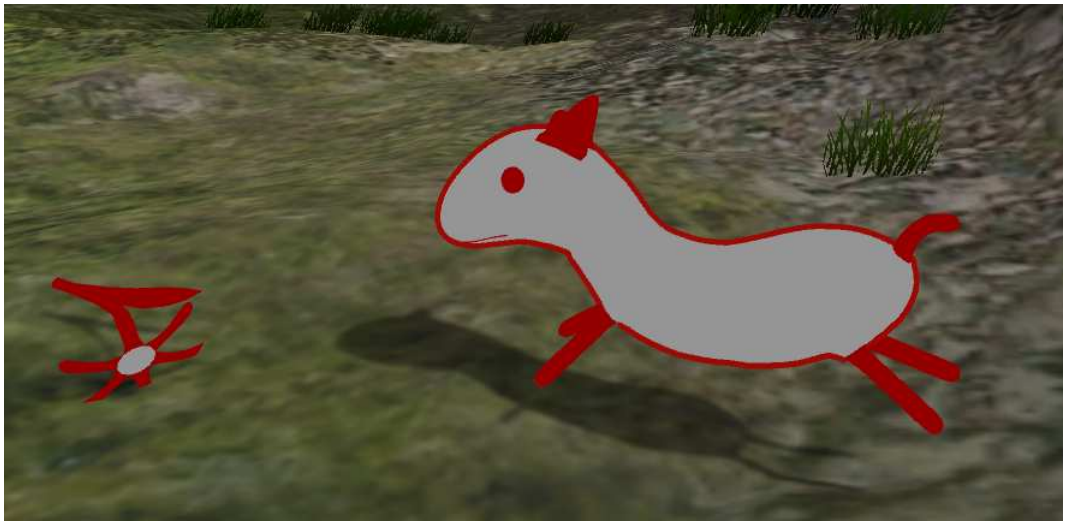


Figure 12.2: Reeya, my virtual pet, recreated with the Unity GameEngine.

Part V

Supplemental Information

Bibliography

[1] Ali E. Abbas. *Foundations of Multiattribute Utility*. Cambridge University Press, 2018. [cited at p. 8, 58, 68, 74, 75, 92, 123, 124, 174, 199, 207]

[2] Elisabeth André, Martin Klesen, Patrick Gebhard, Steve Allen, and Thomas Rist. *Integrating Models of Personality and Emotions into Lifelike Characters*, pages 150–165. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000. [cited at p. 107]

[3] Sean Andrist, Bilge Mutlu, and Adriana Tapus. Look Like Me: Matching Robot Personality via Gaze to Increase Motivation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3603–3612, New York, NY, USA, 2015. ACM. [cited at p. 7, 111, 112, 115, 116, 333]

[4] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational Gaze Aversion for Humanlike Robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, HRI '14, pages 25–32, New York, NY, USA, 2014. ACM. [cited at p. 15, 104, 105, 106, 126, 130, 131, 143, 200, 203, 205, 212, 329, 333]

[5] Diana Arellano, Francisco J. Perales, and Javier Varona. Mood and Its Mapping onto Facial Expressions. In Francisco José Perales and José Santos-Victor, editors, *Articulated Motion and Deformable Objects*, volume 8563 of *Lecture Notes in Computer Science*, pages 31–40. Springer International Publishing, 2014. [cited at p. 263]

[6] Diana Arellano, Javier Varona, Francisco J. Perales, Nikolaus Bee, Kathrin Janowski, and Elisabeth André. Influence of head orientation in perception of personality traits in virtual agents. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '11, pages 1093–1094, Richland, SC, 2011. International Foundation for Autonomous Agents and Multiagent Systems. [cited at p. 108]

[7] Michael Argyle. *Bodily Communication*. Routledge, 2 edition, March 1988. [cited at p. 40, 61, 62]

[8] Michael Argyle and Mark Cook. *Gaze and mutual gaze*. Cambridge University Press, The Pitt Building, Trumpington Street, Cambridge CB2 1RP, 1976. [cited at p. 23, 55, 59, 60, 61, 128, 130, 211]

[9] Michael Argyle and Janet Dean. Eye-Contact, Distance and Affiliation. *Sociometry*, 28(3):289–304, 1965. [cited at p. 61]

[10] Michael Argyle and Jean Ann Graham. The central Europe experiment: Looking at persons and looking at objects. *Environmental psychology and nonverbal behavior*, 1(1):6–16, 1976. [cited at p. 59, 60, 128]

[11] Michael Argyle and Brian R. Little. Do Personality Traits Apply to Social Behaviour? *Journal for the Theory of Social Behaviour*, 2(1):1–33, 1972. [cited at p. 20, 21, 23, 40, 57, 111, 126, 207]

[12] Agnes Axelsson and Gabriel Skantze. Do you follow? A fully automated system for adaptive robot presenters. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, page 102–111, New York, NY, USA, 2023. Association for Computing Machinery. [cited at p. 87]

[13] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. [cited at p. 222, 262]

[14] Gene Ball and Jack Breese. Relating Personality and Behavior: Posture and Gestures. In Ana Paiva, editor, *Affective Interactions*, volume 1814 of *Lecture Notes in Computer Science*, pages 196–203. Springer Berlin Heidelberg, 2000. [cited at p. 9, 21, 107, 117, 124]

[15] John E. Barbuto, Jr and Richard W. Scholl. Motivation Sources Inventory: Development and validation of new scales to measure an integrative taxonomy of motivation. *Psychological Reports - PSYCHOL REP*, 82, 1998. [cited at p. 15, 43]

[16] Tobias Baur, Dominik Schiller, and Elisabeth André. *Modeling User's Social Attitude in a Conversational System*, pages 181–199. Springer International Publishing, Cham, 2016. [cited at p. 104]

[17] Dan Bohus and Eric Horvitz. Decisions about turns in multiparty conversation: From perception to action. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, pages 153–160, New York, NY, USA, 2011. ACM. [cited at p. 8, 9, 98, 107, 174, 200, 241, 243, 256, 258]

[18] Dan Bohus and Eric Horvitz. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proceedings of the SIGDIAL 2011 Conference*, SIGDIAL '11, pages 98–109, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. [cited at p. 9, 14, 98, 107, 174, 175, 241, 258]

[19] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–11, New York, NY, USA, 2019. Association for Computing Machinery. event-place: Glasgow, Scotland Uk. [cited at p. 113, 116]

[20] Penelope Brown and Stephen C. Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, February 1987. Google-Books-ID: OG7W8yA2XjcC. [cited at p. 21, 22, 48, 57, 58, 62, 172, 208, 248]

[21] Hennie Brugman and Albert Russel. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). [cited at p. 228]

[22] Harry Bunt. Guidelines for using ISO standard 24617-2, 2017. [cited at p. 19, 92, 249]

[23] Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, and others. Towards an ISO standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2010. [cited at p. 19, 216]

[24] Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R. Traum. ISO 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pages 430–437, 2012. [cited at p. 141, 216]

[25] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. The NoXi Database: Multimodal Recordings of Mediated Novice-expert Interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017, pages 350–359, New York, NY, USA, 2017. ACM. [cited at p. 101]

[26] Crystal Chao. *Timing multimodal turn-taking in human-robot cooperative activity*. PhD thesis, Georgia Institute of Technology, 2015. [cited at p. 95, 97, 106, 117, 129, 174, 181, 200, 288]

[27] Crystal Chao and Andrea Thomaz. Controlling Social Dynamics with a Parametrized Model of Floor Regulation. *Journal of Human-Robot Interaction*, 2(1):4–29, March 2013. [cited at p. 95, 175, 208]

[28] Mathieu Chollet, Magalie Ochs, and Catherine Pelachaud. From non-verbal signals sequence mining to Bayesian networks for interpersonal attitudes expression. In Timothy Bickmore, Stacy Marsella, and Candace Sidner, editors, *Intelligent Virtual Agents*, pages 120–133, Cham, 2014. Springer International Publishing. [cited at p. 109, 117]

[29] Ada S. Chulef, Stephen J. Read, and David A. Walsh. A Hierarchical Taxonomy of Human Goals. *Motivation and Emotion*, 25(3):191–232, September 2001. [cited at p. 7, 15, 43, 44, 45, 46, 47, 51, 52, 172, 248, 327, 328, 333]

[30] Adam Chýlek, Jan Švec, and Luboš Šmídl. Learning to Interrupt the User at the Right Time in Incremental Dialogue Systems. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech, and Dialogue*, pages 500–508, Cham, 2018. Springer International Publishing. [cited at p. 9, 100, 101]

[31] Herbert H. Clark. When to start speaking, when to stop, and how. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, 2003. [cited at p. 18]

[32] Herbert H. Clark and Susan E. Brennan. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149, 1991. [cited at p. 23, 54, 58, 59]

[33] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artif. Intell.*, 42(2-3):213–261, 1990. [cited at p. 14, 16]

[34] Philip R. Cohen and Hector J. Levesque. Communicative actions for artificial agents. In Victor R. Lesser and Les Gasser, editors, *Proceedings of the First International Conference on Multiagent Systems, June 12-14, 1995, San Francisco, California, USA*, pages 65–72. The MIT Press, 1995. [cited at p. 7, 14, 17, 77, 81, 123]

[35] Cristina Conati. Virtual Butler: What Can We Learn from Adaptive User Interfaces? In Robert Trappl, editor, *Your Virtual Butler*, volume 7407 of *Lecture Notes in Computer Science*, pages 29–41. Springer, Berlin, Heidelberg, 2013. [cited at p. 9, 99, 116, 124]

[36] Nigel Crook, Cameron Smith, Marc Cavazza, Stephen Pulman, Roger Moore, and Johan Boye. Handling user interruptions in an embodied conversational agent. In *Proceedings of the AAMAS International Workshop on Interacting with ECAs as Virtual Characters*, page 27 – 33, Toronto, May 2010. [cited at p. 105, 224]

[37] Boele De Raad. Interpersonal lexicon: Structural evidence from two independently constructed verb-based taxonomies. *European Journal of Psychological Assessment*, 15:181–195, 09 1999. [cited at p. 40]

[38] David DeVault, Kenji Sagae, and David Traum. Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse*, 2(1):143–170, 2011. [cited at p. 5, 9, 94, 97, 200, 263]

[39] Colin G. DeYoung, Yanna J. Weisberg, Lena C. Quilty, and Jordan B. Peterson. Unifying the Aspects of the Big Five, the Interpersonal Circumplex, and Trait Affiliation. *Journal of Personality*, 81(5):465–475, 2013. [cited at p. 32, 34, 35, 41, 126, 131, 212, 248, 251, 291]

[40] M. Brent Donnellan and Richard E. Lucas. Age differences in the Big Five across the life span: Evidence from two national samples. *Psychology and aging*, 23(3):558—566, September 2008. [cited at p. 172]

[41] Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283, 1972. [cited at p. 17, 18, 58, 59]

[42] Michael Fleming and Robin Cohen. A utility-based theory of initiative in mixed-initiative systems. In *The IJCAI-01 Workshop on Autonomy, Delegation, and Control: Interacting with Autonomous Agents*, 2001. [cited at p. 174]

[43] Atsushi Fukayama, Takehiko Ohno, Naoki Mukawa, Minako Sawaki, and Norihiro Hagita. Messages Embedded in Gaze of Interface Agents - Impression Management with Agent's Gaze. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pages 41–48, New York, NY, USA, 2002. ACM. [cited at p. 109]

[44] Patrick Gebhard. *Emotionalisierung interaktiver Virtueller Charaktere. Ein mehrschichtiges Computermodell zur Erzeugung und Simulation von Gefühlen in Echtzeit.* PhD thesis, Universität des Saarlandes, Saarbrücken, Nov 2007. [cited at p. 21, 126, 203, 220]

[45] Patrick Gebhard, Tobias Baur, Ionut Damian, Gregor Mehlmann, Johannes Wagner, and Elisabeth André. Exploring Interaction Strategies for Virtual Characters to Induce Stress in Simulated Job Interviews. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, pages 661–668, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems. [cited at p. 114, 263]

[46] Patrick Gebhard, Gregor Mehlmann, and Michael Kipp. Visual Scenemaker - a tool for authoring interactive virtual characters. *Multimodal User Interfaces*, 6(1-2):3–11, 2012. [cited at p. 81, 179]

[47] Nadine Glas, Angelo Cafaro, and Catherine Pelachaud. The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '16, pages 911–920, Richland, SC, 2016. International Foundation for Autonomous Agents and Multiagent Systems. [cited at p. 21, 110, 169, 174, 175, 178, 188, 198, 208]

[48] Julia A. Goldberg. Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power- and rapport-oriented acts. *Journal of Pragmatics*, 14(6):883 – 903, 1990. [cited at p. 18, 59, 62, 76, 127]

[49] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann Jr. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6):504 – 528, 2003. [cited at p. 28, 29, 30, 31, 32, 115]

[50] Stephan Hammer, Birgit Lugrin, Sergey Bogomolov, Kathrin Janowski, and Elisabeth André. *Investigating Politeness Strategies and Their Persuasiveness for a Robotic Elderly Assistant*, pages 315–326. Springer International Publishing, Cham, 2016. [cited at p. 324]

[51] Arno Hartholt, David Traum, Stacy C. Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. All together now: Introducing the virtual human toolkit. In Ruth Aylett, Brigitte Krenn, Catherine Pelachaud, and Hiroshi Shimodaira, editors, *Intelligent Virtual Agents*, pages 368–381, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. [cited at p. 14, 80]

[52] Fritz Heider and Marianne Simmel. An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, 57(2):243–259, 1944. Publisher: University of Illinois Press. [cited at p. 4]

[53] Willem K. Hofstee, Boele De Raad, and Lewis R. Goldberg. Integration of the Big Five and circumplex approaches to trait structure. *Journal of personality and social psychology*, 63(1):146, 1992. [cited at p. 24, 36, 37, 38, 39, 327]

[54] Aaron Holroyd, Charles Rich, Candace L. Sidner, and Brett Ponsler. Generating connection events for human-robot collaboration. In *RO-MAN, 2011 IEEE*, pages 241–246, 2011. [cited at p. 79, 96, 100, 103, 107]

[55] Leonard M. Horowitz, Kelly R. Wilson, Bulent Turan, Pavel Zolotsev, Michael J. Constantino, and Lynne Henderson. How Interpersonal Motives Clarify the Meaning of Interpersonal Behavior: A Revised Circumplex Model. *Personality and Social Psychology Review*, 10(1):67–86, 2006. [cited at p. 32, 33, 34, 35, 47, 116]

[56] Eric Horvitz, Paul Koch, and Johnson Apacible. BusyBody: creating and fielding personalized models of the cost of interruption. In *Proceedings of the*

2004 ACM conference on Computer supported cooperative work, CSCW '04, pages 507–510, New York, NY, USA, 2004. ACM. [cited at p. 8, 74, 98, 124, 174]

[57] Chien-Ming Huang and Bilge Mutlu. Robot Behavior Toolkit: Generating Effective Social Behaviors For Robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 25–32. ACM, 2012. [cited at p. 79]

[58] Chien-Ming Huang and A.L. Thomaz. Effects of responding to, initiating and ensuring joint attention in human-robot interaction. In *RO-MAN, 2011 IEEE*, pages 65–71, 2011. [cited at p. 103, 126]

[59] Markus Häring, Jessica Eichberg, and Elisabeth André. Studies on grounding with gaze and pointing gestures in human-robot-interaction. *Social Robotics*, pages 378–387, 2012. [cited at p. 103, 159]

[60] ISO/DIS 24617-2. Language resource management – Semantic annotation framework (SemAF) – Part 2: Dialogue Acts, 2010. Draft International Standard, Reference Number ISO/DIS 24617-2(E). [cited at p. 19, 249, 319, 320, 335]

[61] Kathrin Janowski. Künstliche Höflichkeit und Frechheit. Wie erhält ein Pflegeroboter das passende Auftreten? In Willibald J. Stronegger and Johann Platzer, editors, *Technisierung der Pflege: 4. Goldegger Dialogforum Mensch und Endlichkeit*, pages 79–90. Nomos Verlagsgesellschaft mbH & Co. KG, Baden-Baden, 2022. [cited at p. 326]

[62] Kathrin Janowski and Elisabeth André. Deciding when to react to incremental user input in human-robot interaction, 2014. [cited at p. 325]

[63] Kathrin Janowski and Elisabeth André. Decision-theoretic personality-based reasoning about turn-taking conflicts. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA '18, pages 349–350, New York, NY, USA, 2018. ACM. [cited at p. 162, 170, 325]

[64] Kathrin Janowski and Elisabeth André. What If I Speak Now? A Decision-Theoretic Approach to Personality-Based Turn-Taking. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pages 1051–1059, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. [cited at p. 162, 170, 208, 251, 252, 325]

[65] Kathrin Janowski and Elisabeth André. Nichtverbales Verhalten sozialer Roboter. In Oliver Bendel, editor, *Soziale Roboter: Technikwissenschaftliche, wirtschaftswissenschaftliche, philosophische, psychologische und soziologische Grundlagen*, pages 293–308. Springer Fachmedien Wiesbaden, Wiesbaden, 2021. [cited at p. 324]

[66] Kathrin Janowski, Hannes Ritschel, and Elisabeth André. Adaptive artificial personalities. In Birgit Lugrin, Catherine Pelachaud, and David Traum, editors, *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*, volume 2, pages 155–194. Association for Computing Machinery, New York, NY, USA, 2022. [cited at p. 265, 326]

[67] Kathrin Janowski, Hannes Ritschel, Birgit Lugrin, and Elisabeth André. *Sozial interagierende Roboter in der Pflege*, pages 63–87. Springer Fachmedien Wiesbaden, Wiesbaden, 2018. [cited at p. 323]

[68] Mathieu Jégou. *Coordination des tours de parole par le couplage sensorimoteur continu entre utilisateurs et agents*. Phd thesis, Université de Bretagne occidentale - Brest, October 2016. [cited at p. 80]

[69] Mathieu Jégou and Pierre Chevaillier. A computational model for the emergence of turn-taking behaviors in user-agent interactions. *Journal on Multimodal User Interfaces*, 12(3):199–223, September 2018. [cited at p. 80, 256]

[70] Adam Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22 – 63, 1967. [cited at p. 17, 18, 60]

[71] Mark L. Knapp, Judith A. Hall, and Terrence G. Horgan. *Nonverbal Communication in Human Interaction (International Edition)*. Wadsworth, Cengage Learning, 8 edition, 2014. [cited at p. 18, 40, 54, 55, 59, 60, 61, 62]

[72] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. Thórisson, and Hannes Vilhjálmsson. Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. In Jonathan Gratch, Michael Young, Ruth Aylett, Daniel Ballin, and Patrick Olivier, editors, *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science*, pages 205–217. Springer Berlin Heidelberg, 2006. [cited at p. 79, 92]

[73] Stefan Kopp, Herwin van Welbergen, Ramin Yaghoubzadeh, and Hendrik Buschmeier. An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing. *Journal on Multimodal User Interfaces*, 8(1):97–108, March 2014. [cited at p. 80, 92, 154]

[74] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature*, 10(1096), 03 2019. [cited at p. 6, 247]

[75] Quoc Anh Le, Jing Huang, and Catherine Pelachaud. A common gesture and speech production framework for virtual and physical agents. In *ACM international conference on multimodal interaction*, 2012. [cited at p. 80]

[76] Quoc Anh Le and Catherine Pelachaud. Generating Co-speech Gestures for the Humanoid Robot NAO through BML. In Eleni Efthimiou, Georgios Kouroupetroglou, and Stavroula-Evita Fotinea, editors, *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, volume 7206 of *Lecture Notes in Computer Science*, pages 228–237. Springer Berlin Heidelberg, 2012. [cited at p. 15, 79, 80]

[77] Iolanda Leite, André Pereira, Ginevra Castellano, Samuel Mascarenhas, Carlos Martinho, and Ana Paiva. Modelling empathy in social robotic companions. In Liliana Ardissono and Tsvi Kuflik, editors, *Advances in User Modeling*, pages 135–147. Springer Berlin Heidelberg, 2012. [cited at p. 14, 116]

[78] Bertram F. Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 117–124. ACM Press, 2015. [cited at p. 102, 124]

[79] Patrick M. Markey and Charlotte N. Markey. A Brief Assessment of the Interpersonal Circumplex: The IPIP-IPC. *Assessment*, 16(4):352–361, 2009. [cited at p. 32, 33, 34, 35, 191, 212, 251, 291]

[80] Robert R. McCrae and Paul T. Costa. The structure of interpersonal traits: Wiggins's circumplex and the Five-Factor Model. *Journal of personality and social psychology*, 56(4):586, 1989. [cited at p. 20, 21, 24, 32, 33, 34, 35, 126, 131, 212, 251, 291]

[81] Robert R. McCrae and Oliver P. John. An introduction to the Five-Factor Model and its applications. *Journal of personality*, 60(2):175, 1992. [cited at p. 28, 29, 30, 31, 32, 33]

[82] Benjamin Mehlman. Similarity in friendships. *Journal of Social Psychology*, 57(1):195, June 1962. ISBN: 0022-4545. [cited at p. 111]

[83] Gregor Mehlmann, Markus Häring, Kathrin Janowski, Tobias Baur, Patrick Gebhard, and Elisabeth André. Exploring a model of gaze for grounding in multimodal HRI. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 247–254, New York, NY, USA, 2014. ACM. [cited at p. 15, 100, 144, 158, 179, 185, 203, 216, 324]

[84] Gregor Mehlmann, Kathrin Janowski, and Elisabeth André. Modeling grounding for interactive social companions. *KI - Künstliche Intelligenz*, pages 1–8, 09 2015. [cited at p. 144, 158, 179, 185, 216, 324]

[85] Gregor Mehlmann, Kathrin Janowski, Tobias Baur, Markus Häring, Elisabeth André, and Patrick Gebhard. Modeling gaze mechanisms for grounding in HRI. In *Proceedings of the Twenty-first European Conference on Artificial*

Intelligence, ECAI'14, pages 1069–1070, Amsterdam, The Netherlands, The Netherlands, 2014. IOS Press. [cited at p. 158, 203, 324]

[86] Albert Mehrabian. Analysis of the Big-Five Personality Factors in Terms of the PAD Temperament Model. *Australian Journal of Psychology*, 48(2):86–92, August 1996. [cited at p. 24, 25, 28, 29, 30, 31, 32, 36, 40, 41, 42, 327]

[87] Albert Mehrabian. Pleasure-Arousal-Dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology*, 14(4):261–292, December 1996. [cited at p. 20, 23, 24, 25, 26, 27, 28, 57]

[88] Thilo Michael. RETICO: An incremental framework for spoken dialogue systems. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 49–52, 1st virtual meeting, July 2020. Association for Computational Linguistics. [cited at p. 222, 251, 262]

[89] Youngme Moon. Personalization and Personality: Some Effects of Customizing Message Style Based on Consumer Personality. *Journal of Consumer Psychology*, 12(4):313 – 325, 2002. [cited at p. 7, 113, 115, 116]

[90] Fabrizio Morbini, David DeVault, Kenji Sagae, Jillian Gerten, Angela Nazarian, and David Traum. FLoReS: a forward looking, reward seeking, dialogue manager. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 313–325. Springer, 2014. [cited at p. 81]

[91] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. The Uncanny Valley. *IEEE Robotics & Automation Magazine*, 19(2):98–100, June 2012. [cited at p. 288]

[92] Theresa B. Moyers, Lauren N. Rowell, Jennifer K. Manuel, Denise Ernst, and Jon M. Houck. The Motivational Interviewing Treatment Integrity Code (MITI 4): Rationale, preliminary reliability and validity. *Journal of substance abuse treatment*, 65:36–42, 2016. [cited at p. 263]

[93] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Guanhua Zhang, Patrick Gebhard, Elisabeth André, and Andreas Bulling. MultiMediate: Multi-modal group behaviour analysis for artificial mediation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4878–4882, New York, NY, USA, 2021. Association for Computing Machinery. [cited at p. 106]

[94] Richard E. Neapolitan. *Learning Bayesian networks*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, Upper Saddle River, NJ, 2004. OCLC: ocm52534097. [cited at p. 6, 68, 74, 76, 92, 256]

[95] Radoslaw Niewiadomski, Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. Greta: An interactive expressive ECA system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '09, page 1399–1400, Richland, SC, 2009. International Foundation for Autonomous Agents and Multiagent Systems. [cited at p. 14, 80]

[96] Sergei Nirenburg. *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, volume 8 of *Natural Language Processing*, chapter The Maryland virtual patient as a task-oriented conversational Companion, pages 340–373. John Benjamins Publishing, 2010. [cited at p. 5, 263]

[97] Jonathan Oakman, Shannon Gifford, and Natasha Chlebowsky. A Multilevel Analysis of the Interpersonal Behavior of Socially Anxious People. *Journal of Personality*, 71(3):397–434, 2003. [cited at p. 50]

[98] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure Of Emotions*. Cambridge University Press, 2 edition, 2022. [cited at p. 55, 56, 57, 63, 99, 108, 203, 207, 248]

[99] Charles E. Osgood. Interpersonal Verbs and Interpersonal Behavior. Technical Report, ILLINOIS UIV AT URBANA GROUP EFFECTIVENESS RESEARCH LAB, November 1968. [cited at p. 21, 24, 25, 27, 28]

[100] Catherine Pelachaud. Multimodal expressive embodied conversational agents. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, page 683–689, New York, NY, USA, 2005. Association for Computing Machinery. [cited at p. 80]

[101] Volha Petukhova and Harry Bunt. The coding and annotation of multimodal dialogue acts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1293–1300, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). [cited at p. 19, 92, 249]

[102] Colin Pollock, Nikolaus Bee, Elisabeth André, and Marilyn Walker. Bossy or Wimpy: Expressing Social Dominance by Combining Gaze and Linguistic Behaviors. In Jan Allbeck, Norman Badler, Timothy Bickmore, Catherine Pelachaud, and Alla Safonova, editors, *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings*, pages 265–271. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. [cited at p. 5]

[103] Alexandru Popescu, Joost Broekens, and Maarten van Someren. GAMYGDALA: An Emotion Engine for Games. *Affective Computing, IEEE Transactions on*, 5(1):32–44, January 2014. [cited at p. 203]

[104] Beatrice Rammstedt and Oliver P. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1):203 – 212, 2007. [cited at p. 28, 29, 30, 31, 32, 191]

[105] Anand Srinivasa Rao and Michael P. Georgeff. BDI agents: From theory to practice. In *International Conference on Multiagent Systems*, volume 95, 1995. [cited at p. 14, 77]

[106] Brian Ravenet, Angelo Cafaro, Beatrice Biancardi, Magalie Ochs, and Catherine Pelachaud. Conversational Behavior Reflecting Interpersonal Attitudes in Small Group Interactions. In Willem-Paul Brinkman, Joost Broekens, and Dirk Heylen, editors, *Intelligent Virtual Agents*, volume 9238, pages 375–388. Springer International Publishing, Cham, 2015. [cited at p. 21, 110, 127]

[107] Byron Reeves and Clifford Nass. *The Media Equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press, New York, NY, USA, 1996. [cited at p. 4]

[108] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L. Sidner. Recognizing engagement in human-robot interaction. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, pages 375–382. IEEE, 2010. [cited at p. 96, 99, 103, 104, 175, 229]

[109] Charles Rich and Candace L. Sidner. Using Collaborative Discourse Theory to Partially Automate Dialogue Tree Authoring. In Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker, editors, *Intelligent Virtual Agents*, volume 7502 of *Lecture Notes in Computer Science*, pages 327–340. Springer Berlin Heidelberg, 2012. [cited at p. 81]

[110] Hannes Ritschel, Tobias Baur, and Elisabeth André. Adapting a Robot's linguistic style based on socially-aware reinforcement learning. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 378–384, August 2017. ISSN: 1944-9437. [cited at p. 116, 263]

[111] Hannes Ritschel, Kathrin Janowski, Andreas Seiderer, and Elisabeth André. Towards a Robotic Dietitian with Adaptive Linguistic Style. In *European Conference on Ambient Intelligence*. CEUR Workshop Proceedings, November 2019. [cited at p. 326]

[112] Hannes Ritschel, Kathrin Janowski, Andreas Seiderer, Stefan Wagner, and Elisabeth André. Insights on Usability and User Feedback for an Assistive Robotic Health Companion with Adaptive Linguistic Style. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '19, pages 319–320, New York, NY, USA, 2019. ACM. event-place: Rhodes, Greece. [cited at p. 326]

[113] William T. Rogers and Stanley S. Jones. Effects Of Dominance Tendencies On Floor Holding And Interruption Behavior In Dyadic Interaction. *Human Communication Research*, 1(2):113–122, 1975. [cited at p. 17, 62, 169]

[114] James Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11:273–294, 09 1977. [cited at p. 25, 26, 27, 28, 327]

[115] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. [cited at p. 24, 25, 26, 27, 327]

[116] Steven Sabat. Turn-Taking, Turn-Giving, and Alzheimer's Disease. *Georgetown Journal of Languages and Linguistics*, 2, 1991. [cited at p. 5]

[117] Gerard Saucier. Benchmarks: Integrating affective and interpersonal circles with the Big-Five personality factors. *Journal of Personality and Social Psychology*, 62(6):1025, 1992. [cited at p. 24, 36, 37, 38, 39, 40, 327]

[118] Dominik Schiller, Katharina Weitz, Kathrin Janowski, and Elisabeth André. Human-inspired socially-aware interfaces. In Carlos Martín-Vide, Geoffrey Pond, and Miguel A. Vega-Rodríguez, editors, *Theory and Practice of Natural Computing*, pages 41–53, Cham, 2019. Springer International Publishing. [cited at p. 323]

[119] Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams. Continuously predicting and processing barge-in during a live spoken dialogue task. In *Proceedings of the SIGDIAL 2013 Conference*, pages 384–393. Association for Computational Linguistics (ACL), 2013. [cited at p. 105]

[120] Catherine T. Shea, Erin K. Davisson, and Gráinne M. Fitzsimons. Riding Other People's Coattails: Individuals With Low Self-Control Value Self-Control in Other People. *Psychological Science*, 24(6):1031–1036, 2013. _eprint: https://doi.org/10.1177/0956797612464890. [cited at p. 116]

[121] Gabriel Skantze. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67:101178, 2021. [cited at p. 5, 98, 143, 229, 256]

[122] Gabriel Skantze, Anna Hjalmarsson, and Catharine Oertel. Turn-taking, feedback and joint attention in situated human-robot interaction. *Speech Communication*, 65:50 – 66, 2014. [cited at p. 15, 95, 103, 104, 126, 200, 203, 205]

[123] Paul Slovic, Baruch Fischhoff, and Sarah Lichtenstein. Behavioral decision theory. *Annual review of psychology*, 28(1):1–39, 1977. Publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA. [cited at p. 58, 68, 75, 123, 124]

[124] Christopher J. Soto, Oliver P. John, Samuel D. Gosling, and Jeff Potter. Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of personality and social psychology*, 100(2):330—348, February 2011. [cited at p. 172]

[125] Helen Spencer-Oatey. Reconsidering power and distance. *Journal of Pragmatics*, 26(1):1 – 24, 1996. [cited at p. 21, 35, 36, 56, 172]

[126] Helen Spencer-Oatey. (Im)Politeness, Face and Perceptions of Rapport: Unpackaging their Bases and Interrelationships. *Journal of Politeness Research. Language, Behaviour, Culture*, 1(1), January 2005. [cited at p. 57, 58, 248]

[127] Maria Staudte and Matthew W. Crocker. Visual Attention in Spoken Human-robot Interaction. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, HRI '09, pages 77–84, New York, NY, USA, 2009. ACM. [cited at p. 103]

[128] Valentin Taillandier, Dieuwke Hupkes, Benoît Sagot, Emmanuel Dupoux, and Paul Michel. Neural agents struggle to take turns in bidirectional emergent communication. In *The Eleventh International Conference on Learning Representations*, 2023. [cited at p. 101]

[129] Jennifer R. Talevich, Stephen J. Read, David A. Walsh, Ravi Iyer, and Gurveen Chopra. Toward a comprehensive taxonomy of human motives. *PloS one*, 12(2):e0172279–e0172279, February 2017. [cited at p. 7, 15, 16, 47, 48, 49, 50, 51, 53, 54, 248, 328]

[130] Mark ter Maat, Khiet Phuong Truong, and Dirk K. J. Heylen. How Agents' Turn-Taking Strategies Influence Impressions and Response Behaviors. *Presence: Teleoperators and Virtual Environments*, 20(5):412–430, October 2011. [cited at p. 21, 108, 169, 174, 175, 178, 188, 198, 208]

[131] Robert P. Tett and Patrick J. Murphy. Personality and Situations in Co-worker Preference: Similarity and Complementarity in Worker Compatibility. *Journal of Business and Psychology*, 17(2):223–243, December 2002. [cited at p. 111]

[132] Marcus Thiebaux, Stacy Marsella, Andrew Marshall, and Marcelo Kallmann. SmartBody: Behavior realization for embodied conversational agents. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, volume 1, pages 151–158, 01 2008. [cited at p. 80]

[133] Charles Threlkeld, Muhammad Umair, and Jp de Ruiter. Using transition duration to improve turn-taking in conversational agents. In Oliver Lemon, Dilek Hakkani-Tur, Junyi Jessy Li, Arash Ashrafzadeh, Daniel Hernández Garcia, Malihe Alikhani, David Vandyke, and Ondřej Dušek, editors, *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 193–203, Edinburgh, UK, September 2022. Association for Computational Linguistics. [cited at p. 98]

[134] Paul D. Trapnell and Jerry S. Wiggins. Extension of the Interpersonal Adjective Scales to include the Big Five dimensions of personality. *Journal of Personality and Social Psychology*, 59(4):781–790, 1990. [cited at p. 24, 28, 29, 30, 31, 34]

[135] David Traum, David DeVault, Jina Lee, Zhiyang Wang, and Stacy Marsella. Incremental Dialogue Understanding and Feedback for Multiparty, Multimodal Conversation. In Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker, editors, *Intelligent Virtual Agents*, volume 7502 of *Lecture Notes in Computer Science*, pages 275–288. Springer Berlin Heidelberg, 2012. [cited at p. 5, 9, 94, 263]

[136] Herwin van Welbergen, Ramin Yaghoubzadeh, and Stefan Kopp. AsapRealizer 2.0: The next steps in fluent behavior realization for ECAs. In Timothy Bickmore, Stacy Marsella, and Candace Sidner, editors, *Intelligent Virtual Agents*, pages 449–462, Cham, 2014. Springer International Publishing. [cited at p. 80]

[137] Thomas Visser, David Traum, David DeVault, and Rieks op den Akker. Toward a model for incremental grounding in spoken dialogue systems. In *Proceedings of the 12th International Conference on Intelligent Virtual Agents*, 2012. [cited at p. 94, 97, 175, 200, 263]

[138] Jerry S. Wiggins, Paul Trapnell, and Norman Phillips. Psychometric and geometric characteristics of the Revised Interpersonal Adjective Scales (IAS-R). *Multivariate Behavioral Research*, 23(4):517–530, 1988. [cited at p. 21, 32, 33, 34]

[139] Daksitha Senel Withanage Don, Philipp Müller, Fabrizio Nunnari, Elisabeth André, and Patrick Gebhard. ReNeLiB: Real-time neural listening behavior generation for socially interactive agents. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 507–516, New York, NY, USA, 2023. Association for Computing Machinery. [cited at p. 101]

[140] Liu Yang, Catherine Achard, and Catherine Pelachaud. Now or when? Interruption timing prediction in dyadic interaction. In *ACM International Conference on Intelligent Virtual Agents (IVA '23)*. ACM, New York, NY, USA, 9 2023. [cited at p. 101, 102, 264]

[141] Michelle S. M. Yik and James A. Russell. Predicting the Big Two of Affect from the Big Five of Personality. *Journal of Research in Personality*, 35(3):247 – 277, 2001. [cited at p. 24, 40, 41, 42, 327]

[142] Michelle S. M. Yik and James A. Russell. On the relationship between circumplexes: Affect and Wiggins' IAS. *Multivariate Behavioral Research*, 39(2):203–230, 2004. [cited at p. 21]

[143] Qiaoning Zhang, Connor Esterwood, X. Jessie Yang, and Lionel Robert. An Automated Vehicle (AV) like Me? The Impact of Personality Similarities and Differences between Humans and AVs, 2019. [cited at p. 112, 115, 116]

Acronyms

API application programming interface

DOF degrees of freedom

ECA embodied conversational agent

LLM large language model

LSTM long short-term memory

MAU multi-attribute utility

MNI minimum necessary information

NLU natural language understanding

SVM support vector machine

TTS text-to-speech

Glossary

application programming interface A software interface that lets developers access the functionality of a certain device, web service, or proprietary application.

camel case A form of spelling that leaves no space between individual words but capitalizes their initial letters for readability.

degrees of freedom The number of independently controllable joints in a social robot, such as the axes around which its neck and eyes can rotate.

dyad Two parties interacting with each other. They can be humans, artificial agents, or any combination thereof.

embodied conversational agent An agent with either a graphically represented or robotic body. Its main purpose is to communicate with a human user, for example, using natural speech and gestures.

inverse kinematics An animation approach that calculates an agent's joint angles (often iteratively) so that, for example, its arm reaches a given target position or its head, neck, and possibly torso twist to face the given direction.

large language model A language model that was trained on massive datasets of human-autho-red content, such as social media posts, news articles, and literature.

long short-term memory A type of artificial neural network that learns which information has to be kept for later decisions and when certain information can be discarded.

minimum necessary information A term coined by Chao [26] for the part of an utterance that needs to be perceived before a meaningful response is possible. See section 5.2.1 for details.

multi-attribute utility A high-level utility composed of several low-level utilities, for example, using a weighted sum.

persona In user-centered design, a fictional user representing prototypical preferences or requirements for a subset of the target demographic.

prosody The tone of voice and "melody" of the spoken sentence.

reinforcement learning A machine learning approach in which desirable system actions are symbolically rewarded.

support vector machine A machine learning approach that sorts high-dimensional feature vectors into different classes, depending on their location relative to the hyperplane that bisects the multidimensional space.

text-to-speech Synthetic speech audio produced from a given text.

Uncanny Valley A term proposed by Masahiro Mori in 1970, describing the phenomenon that humans feel repulsed by robots that appear almost human. According to this model, likeability increases as an agent becomes more similar to humans but drops sharply before the point where it would be indistinguishable from them. For more information, see the translation of Mori's essay by MacDorman and Kageki in 2012 [91].

Wizard-of-Oz experiment An experimental setup in which a hidden person controls a computer program to simulate the planned functionality.

Appendices

Appendix A

Non-interactive Prototype

A.1 Calculation of the Default Interpersonal Attitude

As explained in section 3.2.3, there is a proven connection between the Interpersonal Circumplex and two of the "Big Five" personality traits, Extraversion and Agreeableness [80, 79, 39]. This relationship can be expressed as follows:

$$Affiliation = cos(\alpha) * Agreeableness - sin(\alpha) * Extraversion$$
$$Status = sin(\alpha) * Agreeableness + cos(\alpha) * Extraversion$$

Possible combinations of the personality trait values were systematically combined to represent the continuous functions as the conditional probabilities for observing discrete outcomes. The rotation angle was chosen as $\alpha = -37.5°$. Tables A.1 through A.10 present the detailed intermediate results. Cells for input and output values are color-coded.

- **red:** very low, [-1.0;-0.6[
- **yellow:** low, [-0.6;-0.2[
- **gray:** neutral, [-0.2;+0.2]
- **green:** high,]+0.2;+0.6]
- **blue:** very high,]+0.6;+1.0]

A.1.1 Status

		Extraversion: very introverted			
		-0.95	-0.85	-0.75	-0.65
Agreeableness: very disagreeable	-0.95	-0.18	-0.10	-0.02	0.06
	-0.85	-0.24	-0.16	-0.08	0.00
	-0.75	-0.30	-0.22	-0.14	-0.06
	-0.65	-0.36	-0.28	-0.20	-0.12
Agreeableness: disagreeable	-0.55	-0.42	-0.34	-0.26	-0.18
	-0.45	-0.48	-0.40	-0.32	-0.24
	-0.35	-0.54	-0.46	-0.38	-0.30
	-0.25	-0.60	-0.52	-0.44	-0.36
Agreeableness: neutral	-0.15	-0.66	-0.58	-0.50	-0.42
	-0.05	-0.72	-0.64	-0.56	-0.49
	0.05	-0.78	-0.70	-0.63	-0.55
	0.15	-0.84	-0.77	-0.69	-0.61
Agreeableness: agreeable	0.25	-0.91	-0.83	-0.75	-0.67
	0.35	-0.97	-0.89	-0.81	-0.73
	0.45	-1.03	-0.95	-0.87	-0.79
	0.55	-1.09	-1.01	-0.93	-0.85
Agreeableness: very agreeable	0.65	-1.15	-1.07	-0.99	-0.91
	0.75	-1.21	-1.13	-1.05	-0.97
	0.85	-1.27	-1.19	-1.11	-1.03
	0.95	-1.33	-1.25	-1.17	-1.09

Table A.1: Uniform sampling of Extraversion and Agreeableness for calculating the Status levels. Table section for Extraversion level *very introverted*.

		Extraversion: introverted			
		-0.55	-0.45	-0.35	-0.25
Agreeableness: very disagreeable	-0.95	0.14	0.22	0.30	0.38
	-0.85	0.08	0.16	0.24	0.32
	-0.75	0.02	0.10	0.18	0.26
	-0.65	-0.04	0.04	0.12	0.20
Agreeableness: disagreeable	-0.55	-0.10	-0.02	0.06	0.14
	-0.45	-0.16	-0.08	0.00	0.08
	-0.35	-0.22	-0.14	-0.06	0.01
	-0.25	-0.28	-0.20	-0.13	-0.05
Agreeableness: neutral	-0.15	-0.35	-0.27	-0.19	-0.11
	-0.05	-0.41	-0.33	-0.25	-0.17
	0.05	-0.47	-0.39	-0.31	-0.23
	0.15	-0.53	-0.45	-0.37	-0.29
Agreeableness: agreeable	0.25	-0.59	-0.51	-0.43	-0.35
	0.35	-0.65	-0.57	-0.49	-0.41
	0.45	-0.71	-0.63	-0.55	-0.47
	0.55	-0.77	-0.69	-0.61	-0.53
Agreeableness: very agreeable	0.65	-0.83	-0.75	-0.67	-0.59
	0.75	-0.89	-0.81	-0.73	-0.65
	0.85	-0.95	-0.87	-0.80	-0.72
	0.95	-1.01	-0.94	-0.86	-0.78

Table A.2: Uniform sampling of Extraversion and Agreeableness for calculating the Status levels. Table section for Extraversion level *introverted*.

		Extraversion: neutral			
		-0.15	-0.05	0.05	0.15
Agreeableness: very disagreeable	-0.95	0.46	0.54	0.62	0.70
	-0.85	0.40	0.48	0.56	0.64
	-0.75	0.34	0.42	0.50	0.58
	-0.65	0.28	0.36	0.44	0.51
Agreeableness: disagreeable	-0.55	0.22	0.30	0.37	0.45
	-0.45	0.15	0.23	0.31	0.39
	-0.35	0.09	0.17	0.25	0.33
	-0.25	0.03	0.11	0.19	0.27
Agreeableness: neutral	-0.15	-0.03	0.05	0.13	0.21
	-0.05	-0.09	-0.01	0.07	0.15
	0.05	-0.15	-0.07	0.01	0.09
	0.15	-0.21	-0.13	-0.05	0.03
Agreeableness: agreeable	0.25	-0.27	-0.19	-0.11	-0.03
	0.35	-0.33	-0.25	-0.17	-0.09
	0.45	-0.39	-0.31	-0.23	-0.15
	0.55	-0.45	-0.37	-0.30	-0.22
Agreeableness: very agreeable	0.65	-0.51	-0.44	-0.36	-0.28
	0.75	-0.58	-0.50	-0.42	-0.34
	0.85	-0.64	-0.56	-0.48	-0.40
	0.95	-0.70	-0.62	-0.54	-0.46

Table A.3: Uniform sampling of Extraversion and Agreeableness for calculating the Status levels. Table section for Extraversion level *neutral*.

		Extraversion: extraverted			
		0.25	0.35	0.45	0.55
Agreeableness: very disagreeable	-0.95	0.78	0.86	0.94	1.01
	-0.85	0.72	0.80	0.87	0.95
	-0.75	0.65	0.73	0.81	0.89
	-0.65	0.59	0.67	0.75	0.83
Agreeableness: disagreeable	-0.55	0.53	0.61	0.69	0.77
	-0.45	0.47	0.55	0.63	0.71
	-0.35	0.41	0.49	0.57	0.65
	-0.25	0.35	0.43	0.51	0.59
Agreeableness: neutral	-0.15	0.29	0.37	0.45	0.53
	-0.05	0.23	0.31	0.39	0.47
	0.05	0.17	0.25	0.33	0.41
	0.15	0.11	0.19	0.27	0.35
Agreeableness: agreeable	0.25	0.05	0.13	0.20	0.28
	0.35	-0.01	0.06	0.14	0.22
	0.45	-0.08	0.00	0.08	0.16
	0.55	-0.14	-0.06	0.02	0.10
Agreeableness: very agreeable	0.65	-0.20	-0.12	-0.04	0.04
	0.75	-0.26	-0.18	-0.10	-0.02
	0.85	-0.32	-0.24	-0.16	-0.08
	0.95	-0.38	-0.30	-0.22	-0.14

Table A.4: Uniform sampling of Extraversion and Agreeableness for calculating the Status levels. Table section for Extraversion level *extraverted*.

		Extraversion: very extraverted			
		0.65	0.75	0.85	0.95
Agreeableness: very disagreeable	-0.95	1.09	1.17	1.25	1.33
	-0.85	1.03	1.11	1.19	1.27
	-0.75	0.97	1.05	1.13	1.21
	-0.65	0.91	0.99	1.07	1.15
Agreeableness: disagreeable	-0.55	0.85	0.93	1.01	1.09
	-0.45	0.79	0.87	0.95	1.03
	-0.35	0.73	0.81	0.89	0.97
	-0.25	0.67	0.75	0.83	0.91
Agreeableness: neutral	-0.15	0.61	0.69	0.77	0.84
	-0.05	0.55	0.63	0.70	0.78
	0.05	0.49	0.56	0.64	0.72
	0.15	0.42	0.50	0.58	0.66
Agreeableness: agreeable	0.25	0.36	0.44	0.52	0.60
	0.35	0.30	0.38	0.46	0.54
	0.45	0.24	0.32	0.40	0.48
	0.55	0.18	0.26	0.34	0.42
Agreeableness: very agreeable	0.65	0.12	0.20	0.28	0.36
	0.75	0.06	0.14	0.22	0.30
	0.85	0.00	0.08	0.16	0.24
	0.95	-0.06	0.02	0.10	0.18

Table A.5: Uniform sampling of Extraversion and Agreeableness for calculating the Status levels. Table section for Extraversion level *very extraverted*.

A.1.2 Affiliation

		Extraversion: very introverted			
		-0.95	-0.85	-0.75	-0.65
Agreeableness: very disagreeable	-0.95	-1.33	-1.27	-1.21	-1.15
	-0.85	-1.25	-1.19	-1.13	-1.07
	-0.75	-1.17	-1.11	-1.05	-0.99
	-0.65	-1.09	-1.03	-0.97	-0.91
Agreeableness: disagreeable	-0.55	-1.01	-0.95	-0.89	-0.83
	-0.45	-0.94	-0.87	-0.81	-0.75
	-0.35	-0.86	-0.80	-0.73	-0.67
	-0.25	-0.78	-0.72	-0.65	-0.59
Agreeableness: neutral	-0.15	-0.70	-0.64	-0.58	-0.51
	-0.05	-0.62	-0.56	-0.50	-0.44
	0.05	-0.54	-0.48	-0.42	-0.36
	0.15	-0.46	-0.40	-0.34	-0.28
Agreeableness: agreeable	0.25	-0.38	-0.32	-0.26	-0.20
	0.35	-0.30	-0.24	-0.18	-0.12
	0.45	-0.22	-0.16	-0.10	-0.04
	0.55	-0.14	-0.08	-0.02	0.04
Agreeableness: very agreeable	0.65	-0.06	0.00	0.06	0.12
	0.75	0.02	0.08	0.14	0.20
	0.85	0.10	0.16	0.22	0.28
	0.95	0.18	0.24	0.30	0.36

Table A.6: Uniform sampling of Extraversion and Agreeableness for calculating the Affiliation levels. Table section for Extraversion level *very introverted*.

		Extraversion: introverted			
		-0.55	-0.45	-0.35	-0.25
Agreeableness: very disagreeable	-0.95	-1.09	-1.03	-0.97	-0.91
	-0.85	-1.01	-0.95	-0.89	-0.83
	-0.75	-0.93	-0.87	-0.81	-0.75
	-0.65	-0.85	-0.79	-0.73	-0.67
Agreeableness: disagreeable	-0.55	-0.77	-0.71	-0.65	-0.59
	-0.45	-0.69	-0.63	-0.57	-0.51
	-0.35	-0.61	-0.55	-0.49	-0.43
	-0.25	-0.53	-0.47	-0.41	-0.35
Agreeableness: neutral	-0.15	-0.45	-0.39	-0.33	-0.27
	-0.05	-0.37	-0.31	-0.25	-0.19
	0.05	-0.30	-0.23	-0.17	-0.11
	0.15	-0.22	-0.15	-0.09	-0.03
Agreeableness: agreeable	0.25	-0.14	-0.08	-0.01	0.05
	0.35	-0.06	0.00	0.06	0.13
	0.45	0.02	0.08	0.14	0.20
	0.55	0.10	0.16	0.22	0.28
Agreeableness: very agreeable	0.65	0.18	0.24	0.30	0.36
	0.75	0.26	0.32	0.38	0.44
	0.85	0.34	0.40	0.46	0.52
	0.95	0.42	0.48	0.54	0.60

Table A.7: Uniform sampling of Extraversion and Agreeableness for calculating the Affiliation levels. Table section for Extraversion level *introverted*.

		Extraversion: neutral			
		-0.15	-0.05	0.05	0.15
Agreeableness: very disagreeable	-0.95	-0.84	-0.78	-0.72	-0.66
	-0.85	-0.77	-0.70	-0.64	-0.58
	-0.75	-0.69	-0.63	-0.56	-0.50
	-0.65	-0.61	-0.55	-0.49	-0.42
Agreeableness: disagreeable	-0.55	-0.53	-0.47	-0.41	-0.35
	-0.45	-0.45	-0.39	-0.33	-0.27
	-0.35	-0.37	-0.31	-0.25	-0.19
	-0.25	-0.29	-0.23	-0.17	-0.11
Agreeableness: neutral	-0.15	-0.21	-0.15	-0.09	-0.03
	-0.05	-0.13	-0.07	-0.01	0.05
	0.05	-0.05	0.01	0.07	0.13
	0.15	0.03	0.09	0.15	0.21
Agreeableness: agreeable	0.25	0.11	0.17	0.23	0.29
	0.35	0.19	0.25	0.31	0.37
	0.45	0.27	0.33	0.39	0.45
	0.55	0.35	0.41	0.47	0.53
Agreeableness: very agreeable	0.65	0.42	0.49	0.55	0.61
	0.75	0.50	0.56	0.63	0.69
	0.85	0.58	0.64	0.70	0.77
	0.95	0.66	0.72	0.78	0.84

Table A.8: Uniform sampling of Extraversion and Agreeableness for calculating the Affiliation levels. Table section for Extraversion level *neutral*.

		Extraversion: extraverted			
		0.25	0.35	0.45	0.55
Agreeableness: very disagreeable	-0.95	-0.60	-0.54	-0.48	-0.42
	-0.85	-0.52	-0.46	-0.40	-0.34
	-0.75	-0.44	-0.38	-0.32	-0.26
	-0.65	-0.36	-0.30	-0.24	-0.18
Agreeableness: disagreeable	-0.55	-0.28	-0.22	-0.16	-0.10
	-0.45	-0.20	-0.14	-0.08	-0.02
	-0.35	-0.13	-0.06	0.00	0.06
	-0.25	-0.05	0.01	0.08	0.14
Agreeableness: neutral	-0.15	0.03	0.09	0.15	0.22
	-0.05	0.11	0.17	0.23	0.30
	0.05	0.19	0.25	0.31	0.37
	0.15	0.27	0.33	0.39	0.45
Agreeableness: agreeable	0.25	0.35	0.41	0.47	0.53
	0.35	0.43	0.49	0.55	0.61
	0.45	0.51	0.57	0.63	0.69
	0.55	0.59	0.65	0.71	0.77
Agreeableness: very agreeable	0.65	0.67	0.73	0.79	0.85
	0.75	0.75	0.81	0.87	0.93
	0.85	0.83	0.89	0.95	1.01
	0.95	0.91	0.97	1.03	1.09

Table A.9: Uniform sampling of Extraversion and Agreeableness for calculating the Affiliation levels. Table section for Extraversion level *extraverted*.

		Extraversion: very extraverted			
		0.65	0.75	0.85	0.95
Agreeableness: very disagreeable	-0.95	-0.36	-0.30	-0.24	-0.18
	-0.85	-0.28	-0.22	-0.16	-0.10
	-0.75	-0.20	-0.14	-0.08	-0.02
	-0.65	-0.12	-0.06	0.00	0.06
Agreeableness: disagreeable	-0.55	-0.04	0.02	0.08	0.14
	-0.45	0.04	0.10	0.16	0.22
	-0.35	0.12	0.18	0.24	0.30
	-0.25	0.20	0.26	0.32	0.38
Agreeableness: neutral	-0.15	0.28	0.34	0.40	0.46
	-0.05	0.36	0.42	0.48	0.54
	0.05	0.44	0.50	0.56	0.62
	0.15	0.51	0.58	0.64	0.70
Agreeableness: agreeable	0.25	0.59	0.65	0.72	0.78
	0.35	0.67	0.73	0.80	0.86
	0.45	0.75	0.81	0.87	0.94
	0.55	0.83	0.89	0.95	1.01
Agreeableness: very agreeable	0.65	0.91	0.97	1.03	1.09
	0.75	0.99	1.05	1.11	1.17
	0.85	1.07	1.13	1.19	1.25
	0.95	1.15	1.21	1.27	1.33

Table A.10: Uniform sampling of Extraversion and Agreeableness for calculating the Affiliation levels. Table section for Extraversion level *very extraverted*.

A.2 Evaluation

A.2.1 Advertising the Survey

Figure A.1: Design of the flyers and posters distributed across the campus of Augsburg University.

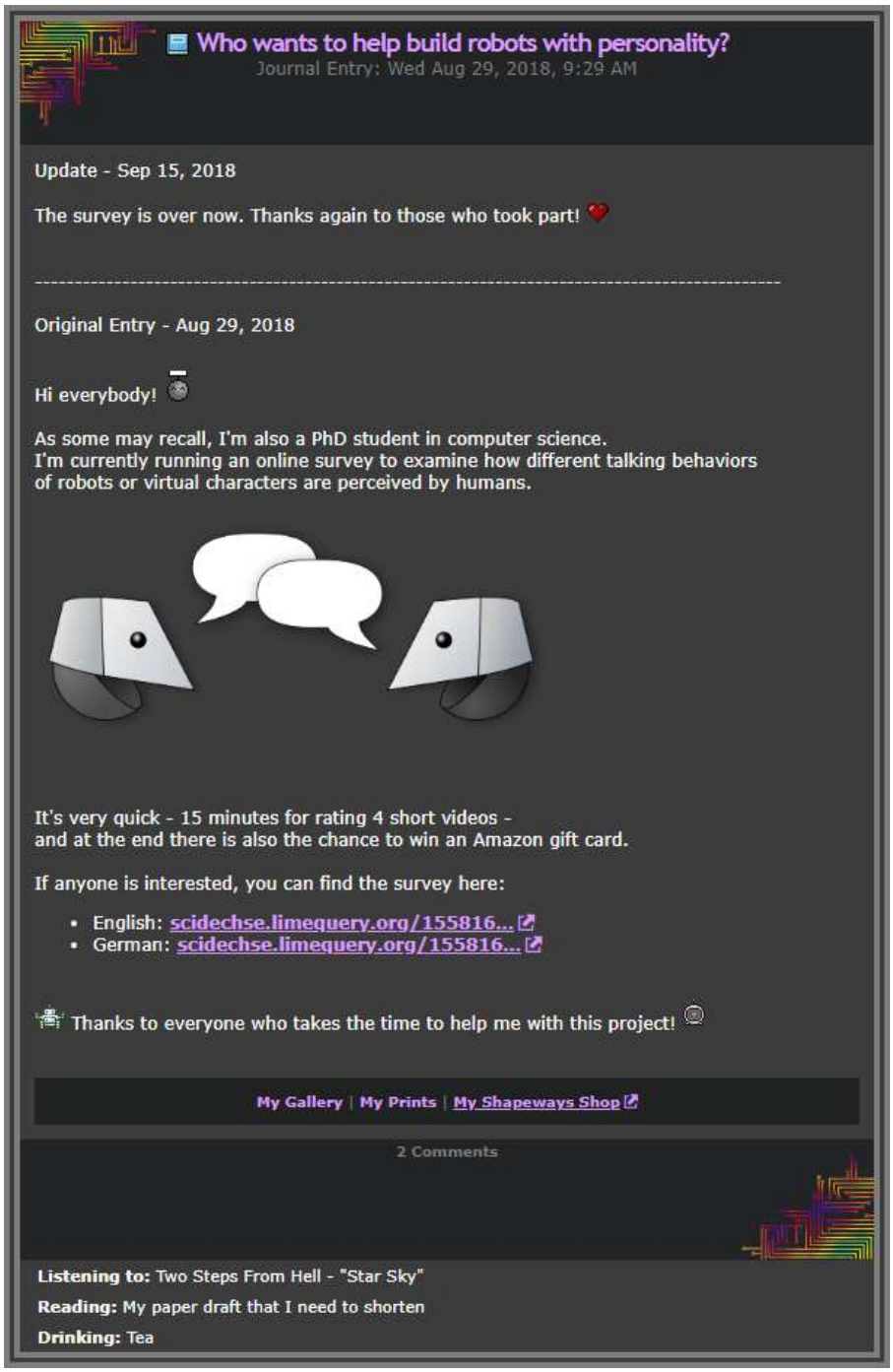


Figure A.2: Journal entry on the artist platform DeviantArt.

A.2.2 Online Survey - German Version

Wahrnehmung von Gesprächsverhalten

Diese Umfrage soll ermitteln, wie das Gesprächsverhalten von virtuellen Charakteren auf menschliche Beobachter wirkt. In dieser Umfrage werden Sie vier kurze Videos sehen, in denen sich zwei virtuelle Charaktere unterhalten. Der Text, den sie sprechen, ist unwichtig und wurde deswegen durch bedeutungslose Laute ersetzt.

Dennoch sollten Sie den Ton einschalten, um besser zu erkennen, wer gerade spricht.

Ihre Aufgabe ist es, deren Sprechverhalten zu beobachten und Ihre Meinung zu deren Persönlichkeit und Beziehung abzugeben.

Bitte antworten Sie spontan, ohne lange nachzudenken. Es gibt keine falschen Antworten.

Die Studie wird etwa 15 Minuten dauern.

Zum Dank für Ihre Teilnahme sind Sie im Anschluss zu einer Verlosung eingeladen. Sie haben die Chance, einen von drei Amazon-Gutscheinen im Wert von jeweils 10€ zu gewinnen.

Über uns: Der Lehrstuhl für Multimodale Mensch-Technik-Interaktion der Universität Augsburg befasst sich in Forschung und Lehre mit der Simulation von menschlichen Verhaltensweisen durch Roboter oder virtuelle Charaktere. Eine wichtige Komponente unserer Arbeit ist die Durchführung von Perzeptionsstudien, um die simulierten Verhaltensweisen zu evaluieren.

Das Projekt: Der Lehrstuhl für Multimodale Mensch-Technik-Interaktion arbeitet derzeit an einem Forschungsprojekt, das sich mit der Simulation von kommunikativen Verhaltensweisen beschäftigt. Diese Studie hat das Ziel, die an unserem Lehrstuhl entwickelten technischen Verfahren durch menschliche Beobachter bewerten zu lassen.

Ansprechpartner: Bei Rückfragen wenden Sie sich bitte per E-Mail an Kathrin Janowski unter folgender Adresse: kathrin.janowski@informatik.uni-augsburg.de

Datenschutzrechtliche Einwilligung

☐ Ich willige ein, dass meine personenbezogenen Daten (demographische Daten, Bewertung der Videos) zum Zweck der Durchführung der Studie zum kommunikativen Verhalten von Robotern und virtuellen Charakteren des Lehrstuhls für Multimodale Mensch-Technik-Interaktion verarbeitet werden dürfen. Darüber hinaus bin ich damit einverstanden, dass meine E-Mail-Adresse für die Zwecke der Verlosung von drei Amazon-Gutscheine und der Benachrichtigung der Gewinner verarbeitet wird.

Vorbereitung

Bevor Sie beginnen, prüfen Sie bitte, ob der Ton gut zu hören ist. Sehen Sie sich bitte dieses Video an und beantworten Sie anschließend die Frage dazu.

sound test video

Wie lautet der vierte Satz, der gesagt wird?

. .....

Allgemeine Informationen

Diese Angaben helfen uns, Ihre Antworten besser zu verstehen.

Alter

- ○ unter 20
- ○ 20 bis 29
- ○ 30 bis 39
- ○ 40 bis 59
- ○ 60 bis 79
- ○ 80 oder darüber
- ○ keine Antwort

Geschlecht

- ○ männlich
- ○ weiblich
- ○ sonstiges: .....
- ○ keine Antwort

Muttersprache

. .....

Mit welcher Sprache sind Sie als Kind aufgewachsen?

Beruf

. .....

Falls zutreffend, geben Sie bitte auch die Fachrichtung an, z.B. "Student (Informatik)" oder "Lehrer (Mathematik, Biologie)".

bisherige Erfahrung mit computer-gesteuerten Charakteren

	überhaupt keine Erfahrung	habe bereits in Aktion gesehen	habe bereits selbst verwendet	verwende regelmäßig selbst
Videospiel-Charaktere	○	○	○	○
Sprachassistenten	○	○	○	○
soziale Roboter	○	○	○	○

- *Videospiel-Charaktere: Figuren im Spiel, mit denen der Spieler sich unterhalten kann, z.B. Auftraggeber oder Weggefährten*
- *Sprachassistenten: sprachgesteuerte Geräte oder Software zur Unterstützung bei Alltagsaufgaben, wie z.B. Siri, Alexa/Amazon Echo, Cortana*
- *soziale Roboter: Roboter, welche in Aussehen und/oder Verhalten Lebewesen ähneln, z.B. NAO und Pepper, Reeti, Jibo*

Video[1]

Bitte sehen Sie sich dieses Video an und beantworten Sie anschließend die untenstehenden Fragen.

stimulus video

Bezogen auf dieses Video, wie sehr stimmen Sie den folgenden Aussagen zu?

Der **linke** Sprecher...

	stimme überhaupt nicht zu	stimme nicht zu	neutral	stimme zu	stimme vollkommen zu
...ist gesprächig.	○	○	○	○	○
...geht aus sich heraus, ist gesellig.	○	○	○	○	○
...ist zurückhaltend, reserviert.	○	○	○	○	○
...ist unhöflich.	○	○	○	○	○
...hat ein versöhnliches Wesen.	○	○	○	○	○
...ist rücksichtsvoll.	○	○	○	○	○
...ist freundlich.	○	○	○	○	○
...ist hat einen niedrigen Rang.	○	○	○	○	○
...kontrolliert das Gespräch.	○	○	○	○	○

[1]These questions are displayed four times, once for each of the stimulus videos.

Der **rechte** Sprecher...

	stimme überhaupt nicht zu	stimme nicht zu	neutral	stimme zu	stimme vollkommen zu
...ist gesprächig.	O	O	O	O	O
...geht aus sich heraus, ist gesellig.	O	O	O	O	O
...ist zurückhaltend, reserviert.	O	O	O	O	O
...ist unhöflich.	O	O	O	O	O
...hat ein versöhnliches Wesen.	O	O	O	O	O
...ist rücksichtsvoll.	O	O	O	O	O
...ist freundlich.	O	O	O	O	O
...ist hat einen niedrigen Rang.	O	O	O	O	O
...kontrolliert das Gespräch.	O	O	O	O	O

Weitere Anmerkungen

. .....

Hier können Sie angeben, was Ihnen sonst noch an dem Video aufgefallen ist.

Verlosung

Zum Dank für Ihre Teilnahme haben Sie jetzt die Chance, einen von drei Amazon-Gutscheinen im Wert von je 10€ zu gewinnen. Falls Sie an der Verlosung teilnehmen wollen, geben Sie bitte Ihre E-Mail-Adresse an. Diese wird ausschließlich für die Verlosung verwendet und danach gelöscht.

E-Mail-Adresse

. .....

Bitte geben Sie eine gültige E-Mail-Adresse an, z.B. "ihr-name@provider.de".

Vielen Dank für Ihre Teilnahme!

Die Gewinner der Verlosung werden in einigen Wochen per E-Mail benachrichtigt.

A.2.3 Online Survey - English Version

Perception of Conversational Behavior

This survey examines how the conversational behavior of virtual characters is perceived by human observers. In this survey you will see four short videos in which two virtual characters are talking to each other. The text which they speak is not important and was therefore replaced with meaningless sounds.

Nevertheless you should turn on the sound to better detect who is speaking.

Your task is to observe their speech behavior and state your opinion on their personality and relationship.

Please answer quickly without thinking too long about it. There are no incorrect answers.

The survey will take about 10-15 minutes.

To thank you for your participation you are invited to enter a lottery afterwards. You'll have the chance to win one of three Amazon gift cards with a value of 10€ each.

About us: The chair of Human-Centered Multimedia at Augsburg University focuses, in research and teaching, on the simulation of human-like behaviors for robots and virtual characters. One important part of our work is the conduction of perception studies to evaluate the simulated behaviors.

The Project: The chair of Human-Centered Multimedia is currently working on a research project which focuses on the simulation of communicative behaviors. The study aims to have human observers evaluate the technical procedures which have been developed at our chair

Contact Person: For further information please contact Kathrin Janowski at the following mail address: kathrin.janowski@informatik.uni-augsburg.de

Data Policy Consent

☐ I grant permission that my personal data (demographic data, rating of the videos) be processed for the purpose of conducting the study about the communicative behavior of robots and virtual characters at the chair of Human-Centered Multimedia. Furthermore I consent to my email being used for the purpose of a lottery for three Amazon gift cards and the notification of the winners.

Preparation

Before you start, please make sure that you can hear the sound well. Please watch this video and answer the question afterwards.

sound test video

What is the fourth sentence which is spoken?

. .....

General Information

This information will help us to better understand your answers.

Age

- ○ under 20
- ○ 20 to 29
- ○ 30 to 39
- ○ 40 to 59
- ○ 60 to 79
- ○ 80 or above
- ○ No answer

Gender

- ○ male
- ○ female
- ○ other: .....
- ○ No answer

First Language

. .....

With which language did you grow up as a child?

Occupation

. .....

If applicable, please add your subject area, e.g. "student (computer science)" or "teacher (mathematics, biology)".

Previous Experience With Computer-Controlled Characters

	no experience at all	have seen it in action	have used it myself	use it regularly myself
Video Game Characters	○	○	○	○
Voice Assistants	○	○	○	○
Social Robots	○	○	○	○

- *Video Game Characters: characters in the game to whom the player can talk, e.g. quest givers or traveling companions*
- *Voice Assistants: speech-controlled devices or software for supporting everyday tasks, e.g. Siri, Alexa/Amazon Echo, Cortana*
- *Social Robots: robots whose appearence and/or behavior resembles living creatures, e.g. NAO and Pepper, Reeti, Jibo*

Video[2]

Please watch this video and then answer the questions below.

stimulus video

With regards to this video, how much do you agree with the following statements?

The speaker on the **left**...

	disagree completely	disagree	neutral	agree	agree completely
...is talkative.	O	O	O	O	O
...is outgoing, sociable.	O	O	O	O	O
...is reserved.	O	O	O	O	O
...is rude.	O	O	O	O	O
...has a forgiving nature.	O	O	O	O	O
...is considerate.	O	O	O	O	O
...is friendly.	O	O	O	O	O
...has a low rank.	O	O	O	O	O
...controls the conversation.	O	O	O	O	O

[2]These questions are displayed four times, once for each of the stimulus videos.

The speaker on the **right**...

	disagree completely	disagree	neutral	agree	agree completely
...is talkative.	O	O	O	O	O
...is outgoing, sociable.	O	O	O	O	O
...is reserved.	O	O	O	O	O
...is rude.	O	O	O	O	O
...has a forgiving nature.	O	O	O	O	O
...is considerate.	O	O	O	O	O
...is friendly.	O	O	O	O	O
...has a low rank.	O	O	O	O	O
...controls the conversation.	O	O	O	O	O

Additional Comments

. .....

Here you can enter anything else that you noticed about the video.

Lottery

As a Thank-You for your participation you now have the chance to win one of three Amazon gift cards with a value of 10€. If you want to take part in the lottery, please enter your mail address. It will solely be used for the lottery and will be deleted afterwards.

Mail Address

. .....

Please enter a valid mail address, e.g. your-name@provider.com

---

Thank you very much for participating!

The lottery winners will be notified by email in a few weeks.

A.2.4 Participant Comments

Control and Initiative

- "Die Linke ist für mich scheinbar bisher immer die Referenz, weil sie das Gespräch beginnt. Die Rechte reagiert ja erstmal nur."
  "The left one always seems to be the reference for me until now because she starts the conversation. The right one is only responding at first."

- "Es hört sich nach einem professionellen Gespräch an, wobei der linke Sprecher etwas berichtet und der rechte die Informationen aufnimmt und Entscheidungen trifft, bzw. Kommentare dazu abgibt."
  "It sounds like a professional conversation, with the left speaker reporting something and the right one receiving the information and making decisions respectively commenting on it."

- "Kontrolle ist schwer zu sagen nur mit Lauten."
  "Control is hard to tell from sounds alone."

Type of Conversation

- "Die Tonlagen erscheinen mir gleich auf beiden Seiten. Ich habe den Eindruck, dass es eher eine Art Streitgespräch ist, wobei die rechte Figur dafür doch zu wenig redet. Aber die Tonlage ist auf beiden Seiten nicht positiv und sie erscheint mir auf beiden Seiten gleich und monoton."
  "To me, the tone of voice seems identical on both sides. I'm under the impression that it's more of a kind of argument, although the right figure does speak too little for that. But the tone of voice on both sides is not positive, and to me, it appears identical and monotonous on both sides."

- "Man kann sich durchaus ständig unterbrechen ohne unhöflich zu sein, beispielsweise, wenn Wissenschaftler untereinander leidenschaftlich diskutieren und sich wechselseitig sagen, was sie erwarten, was der andere gerade sagen will. Das ist aber mit einer bestimmten Gesprächsatmosphäre verbunden. Da der "Singsang" der virtuellen männlichen WASP-Computer genau so monoton ist wie im Video zuvor, liegt hier keine solche simulierte Gesprächsatmosphäre vor."
  "It is certainly possible to interrupt each other without being impolite, for example, when scientists are having a passionate discussion and tell each other what they expect that the other is intending to say. But that is linked to a certain conversational atmosphere. Since the "singsong" of the virtual male WASP computers is exactly as monotonous as in the previous video, no such simulated conversational atmosphere is given here."

- "Bei einem Gespräch zwischen Vorgesetztem und Mitarbeiter wäre es durchaus denkbar, dass der Mitarbeiter zu reden beginnt, bevor der Vorgesetzte ganz fertig gesprochen hat, weil er weiß, was dieser von ihm will. Das ist dann auch nicht unhöflich."
  "In a conversation between supervisor and employee, it is certainly imaginable that the employee starts talking before the supervisor has finished speaking completely because he knows what the latter wants of him. That is not impolite, either."

- "Unterhalten die sich mit einander? Oder sitzen sie im Call-Center und führen jeweils eigene Telefongespräche?"
  "Are they having a conversation with each other? Or are they sitting in the call center, and each one is conducting their own phone conversation?"

Other Aspects

- "Es handelt sich um eine Männerstimme. Es handelt sich um Pseudo-Amerikanisches Englisch. Es wird 'männlich' intoniert, nicht weiblich, obwohl die Computer'masken' geschlechtsneutral sind. Der Sprecher 'ist' ein 40-60 jähriger virtueller männlicher Amerikaner (WASP). Der Sprecher hat also einen gewaltigen Geschlechts-Bias. Das muß bei der Weiterentwicklung berücksichtigt werden.
  "It is a man's voice. It is pseudo-American English. The intonation is 'male', not female, although the computer 'masks' are gender-neutral. The speaker 'is' a 40- to 60-year-old virtual male American (WASP). Therefore, the speaker has an enormous gender bias. This needs to be considered in further development."

- "Sehr spekulativ, die ganzen Merkmale aus dem Video abzuleiten."
  "Very speculative, deducing all these features from the video."

- "Sprechen mir eine nicht bekannte Sprache."
  "Speaking a language unknown to me."

- "Der linke Sprecher spricht teilweise undeutlich."
  "The left speaker speaks incomprehensibly at times."

- "Oft haben die Charaktere weiter den Mund bewegt, obwohl der Ton schon zu Ende war."
  "The characters often kept moving their mouths although the sound was already finished."

Appendix B

Interactive Prototype

B.1 Semantic Feature Structures

Table B.1 maps the communicative acts used in the interactive prototype (see chapter 10) to the 2010 version of the DiAML standard [60]. Note that accept* and decline* are not associated with information about what exactly is accepted or declined. This is because a participant may simply say "yes" or "no", leaving it up to the dialogue manager to associate it with something that was said earlier.

Since most of the communicative acts in that prototype belonged to the *general purpose* dimension, the dimension attribute was omitted from the implementation. The *social obligations management* functions were combined into a single function named "social", with the content specifying the precise function. In a similar vein, the single use of the "autoPositive" function made it appear cumbersome to implement the entire *feedback* dimension.

Another notable difference is the category of *information seeking* functions. While the 2010 DiAML standard distinguishes between different grammatical question types, those acts are treated as *requests* in this thesis. The reason behind that choice was that the dialogue manager was not supposed to do complex reasoning on the exact question types. Instead, most information-seeking acts were going to be simple prompts for one specific fact. In other words, they would request the other participant to provide a particular piece of information.

One function that had no equivalent in the 2010 DiAML standard was the function "transfer". It is used at one single point in the revised "salesperson" scenario when the computer-controlled agent acts as if it were handing over a physical brochure to the human interlocutor.

DiAML		**own implementation**	
dimension	**function**	**function**	**content**
general purpose	question	request	info: *identifier*
	inform	inform	subject: *identifier* property: (*key*:*value*)
			object: *identifier*
			state: *sensor input*
	confirm	confirm	subject: *identifier* property: (*key*: *value*)
	offer	offer	item: *identifier*
	request	request	action: *identifier*
	accept*	accept	—
	acceptOffer	accept	offer: *identifier*
	decline*	decline	—
social obligations management	greeting	social	type: greeting
	goodbye	social	type: goodbye
	thanking	social	type: thanks
feedback	autoPositive	feedback	subject: *identifier*
	—	transfer	item: *identifier*

Table B.1: Mapping between the communicative functions in the DiAML standard [60] (2010 version) and the equivalent communicative acts (function and associated content) used in this thesis. The * is a placeholder meaning "offer", "request" or any other act that is accepted/rejected.

B.2 Analyzing Evaluation Data with R

The respective dataset ("seize vs. wait" and "yield vs. hold") was prepared in the form shown in table B.2.

The following script was used to analyze each dataset after replacing the file name "dataset.csv" with the path to the respective file.

```
1 library(FSA)
2
3 dataset <- read.table("dataset.csv", header=TRUE)
4
5 aggregate(ratio ~ type,  data = dataset,
6     function(x) round(c(mean = mean(x),
7     sd = sd(x)), 5)
8 )
```

type	...	sample	ratio
aggressive	...	aggressive01	0
...	...	...	...
dutiful	...	dutiful03	0.3333333333
...	...	...	...
friendly	...	friendly06	0.6
...	...	...	...
lazy	...	lazy07	0.8333333333

Table B.2: Excerpt of the dataset for $ratio_{yield}$, the relative frequency of the agent yielding the turn when the user talks over it.

```
9
10 results_kruskal <-
11     kruskal.test(ratio ~ type, data = dataset)
12 print(results_kruskal)
13
14 results_dunn <-
15     dunnTest(ratio ~ type, data = dataset, method="holm")
16 print(results_dunn)
```

B.3 Evaluating the Behavior Generation

The following code snippet systematically varies the evidence set in the influence diagram and writes the optimal decisions to a given file.

```
1 //------------------------------------
2 // preparation
3 //------------------------------------
4
5 // reset the whole influence diagram
6 mNetwork.clearAllEvidence();
7
8 // set the evidence for the conversation context
9 for(Entry<String,String> entry: baseObservations.entrySet())
10     mNetwork.setEvidence(entry.getKey(), entry.getValue());
11
12 // set the personality nodes to the first combination
13 for(String node: nodesToVary)
14     mNetwork.setEvidence(node, 0);
15
16 // calculate the number of possible combinations
17 int combinations=1;
```

```
for(String node: nodesToVary)
    combinations = combinations*mNetwork.getOutcomeCount(node);

//-------------------------------------
// testing
//-------------------------------------

for (int i=0; i<combinations; i++){
    // evaluate the current combination
    mNetwork.updateBeliefs();

    // lock the decisions in the correct order
    for(String decisionId: decisionNodes){
    // check: is it a decision node?
    if (nodeType == NodeType.List){
        // find the decision with the
        // highest expected utility
        int bestIdx = getBestOutcome(decisionId);

        // lock that decision
        mNetwork.setEvidence(decisionId, bestIdx);
        mNetwork.updateBeliefs();
    }
}

// log the current combination and the resulting decisions
logCurrentState();

// unlock the decisions again
for(String decision: decisionNodes)
    mNetwork.clearEvidence(decision);

// move forward
setNextCombination();
}
```

Appendix C

Publications During This Thesis

C.1 Applications for Agents with Personality

[67]	**"Sozial interagierende Roboter in der Pflege"**
	K. Janowski, H. Ritschel, B. Lugrin and E. André
	chapter in "Pflegeroboter" (O. Bendel), 2018
	"Was machen soziale Maschinen heute schon?"
	K. Janowski
	invited talk at Careum Dialog 2020
[118]	**"Human-Inspired Socially-Aware Interfaces"**
	D. Schiller, K. Weitz, K. Janowski and E. André
	full paper at "Theory and Practice of Natural Computing", 2019

Table C.1: Publications produced in the context of this thesis, focusing on applications which would benefit from agents with personality.

C.2 Behaviors Related to Communicative Intentions

[85]	**"Modeling Gaze Mechanisms for Grounding in HRI"**
	G. Mehlmann, K. Janowski, T. Baur, M. Häring, E. André and P. Gebhard
	short paper at European Conference on Artificial Intelligence (ECAI) 2014
[83]	**"Exploring a Modeling of Gaze for Grounding in Multimodal HRI"**
	G. Mehlmann, M. Häring, K. Janowski, T. Baur, P. Gebhard and E. André
	full paper at International Conference on Multimodal Interaction (ICMI) 2014
[84]	**"Modeling Grounding for Interactive Social Companions"**
	G. Mehlmann, K. Janowski and E. André
	article in journal "KI - Künstliche Intelligenz" 30/1, 2016
[50]	**"Investigating Politeness Strategies and their Persuasiveness for a Robotic Elderly Assistant"**
	S. Hammer, B. Lugrin, S. Bogomolov, K. Janowski and E. André
	full paper at Persuasive Technology: 11th International Conference (PERSUASIVE) 2016
[65]	**"Nichtverbales Verhalten sozialer Roboter"**
	K. Janowski and E. André
	book chapter in "Soziale Roboter" (O. Bendel), 2021

Table C.2: Publications produced in the context of this thesis, focusing on inferring and expressing communicative intentions via nonverbal behavior.

C.3 Intentions Grounded in Personality and Relationship

[62]	**"Deciding When To React To Incremental User Input In Human-Robot Interaction"**
	K. Janowski and E. André
	workshop paper at International Conference on Human-Robot Interaction (HRI) 2014
[63]	**"Decision-theoretic Personality-based Reasoning about Turn-taking Conflicts"**
	K. Janowski and E. André
	extended abstract at International Conference on Intelligent Virtual Agents (IVA) 2018
[64]	**"What If I Speak Now? A Decision-theoretic Approach to Personality-based Turn-taking"**
	K. Janowski and E. André
	full paper at International Conference on Autonomous Agents and MultiAgent Systems (AAMAS) 2019

Table C.3: Publications produced in the context of this thesis, focusing on relating personality traits and interpersonal attitude to specific interaction goals.

C.4 Matching Agent Behavior To The User

[112]	**"Insights on Usability and User Feedback for an Assistive Robotic Health Companion with Adaptive Linguistic Style"**
	H. Ritschel, K. Janowski, A. Seiderer, S. Wagner and E. André
	short paper at International Conference on PErvasive Technologies Related to Assistive Environments (PETRA) 2019
[111]	**"Towards a Robotic Dietitian with Adaptive Linguistic Style"**
	H. Ritschel, K. Janowski, A. Seiderer and E. André
	short paper at European Conference on Ambient Intelligence (AmI) 2019
[61]	**"Künstliche Höflichkeit und Frechheit. Wie erhält ein Pflegeroboter das passende Auftreten?"**
	K. Janowski
	invited talk and proceedings chapter at "Technisierung der Pflege: 4. Goldegger Dialogforum Mensch und Endlichkeit" 2021
[66]	**"Adaptive Artificial Personalities"**
	K. Janowski, H. Ritschel and E. André
	book chapter in "The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application" (B. Lugrin, C. Pelachaud and D. Traum), 2022

Table C.4: Publications produced in the context of this thesis, focusing on choosing or adapting the agent's behavior to fit the user's requirements.

List of Figures

List of Tables