Serious Role-Playing Games for the Training of Social Skills

Von der Fakultät für Ingenieurwissenschaften

Abteilung Informatik und Angewandte Kognitionswissenschaft

der Universität Duisburg-Essen

Zur Erlangung des akademischen Grades

DOKTOR DER INGENIEURWISSENSCHAFTEN (DR.-ING.)

genehmigte kumulative Dissertation

von

Julia Katharina Othlinghaus-Wulhorst
geb. Othlinghaus
aus
Bonn

1. Gutachter: Prof. Dr. Heinz Ulrich Hoppe
2. Gutachter: Prof. Dr. Andreas Harrer
3. Gutachter: Prof. Dr. German Neubaum

Datum der mündlichen Prüfung: 22.02.2023

ABSTRACT

In recent years, serious games have been established as an efficient medium in education and professional training. They have the capability to be effective tools to promote learning and encourage behavioral change, and they constitute a vital instrument for a variety of education and training scenarios. The combination of the serious gaming approach with role-playing is particularly promising, as authentic simulated environments provide mobile, safe and continuable settings for learners in which they can assume roles in specific contexts, explore new situations, and learn how to act and react without having to fear consequences in the real world.

A special challenge with this kind of games is shaping the pedagogical outcomes, as the effects generally depend on post-role-play reflection. Without feedback and reflection, the transfer to real world situations cannot be ensured. Computer-supported analyses can help to track and evaluate the learners' performances, generate feedback, and provide structured recordings enriched with helpful features like integrated indexing, navigation instruments, search functions, and cross references between different media and data sources.

This thesis focuses on serious role-playing games for the professional training of social skills featuring intelligent support. Besides presenting the broad theoretical foundations, contextual background, and existing research in this field, it proposes a novel conceptual and technical framework for the design and implementation of such games and presents different case studies and evaluation approaches.

From the design perspective, the framework is characterized by chat-like interaction with scripted chatbots in a dialogic setting, a separation of immersion and reflection phases that is considered to be conducive for learning, and computer-generated adaptive feedback based on individual analyses. From the technical point of view, the framework is based on three main components: AI-controlled chatbots that adapt to the learners' behavior, a multi-agent blackboard system to keep components independent and support performance optimization using parallel processing, and intelligent support for automated performance analyses and feedback generation.

This thesis presents two main case studies based on the framework. They showcase two different application fields for serious role-playing games for social skills training: conflict management (multi-user environment) and customer complaint management (single-user environment). Both training scenarios have been evaluated in mixed-method studies, combining different qualitative and quantitative methods of analysis to investigate how learners perceive the behavior of the chatbots and whether the training scenarios qualify as real training situations. Furthermore, the relation between the convergence of visual foci of attention and cooperation quality is analyzed in the collaborative conflict management scenario. The mixed-method approach including both, subjective and objective measures, allows for a complete and synergetic utilization of data and creates a solid foundation for drawing conclusions regarding the research objectives.

LIST OF INCLUDED PUBLICATIONS

Othlinghaus-Wulhorst, J., Hoppe, H. U. (2020). A Technical and Conceptual Framework for Serious Role-Playing Games in the Area of Social Skills Training. *Frontiers in Computer Science*. 2(28), doi: 10.3389/fcomp.2020.00028.

Othlinghaus-Wulhorst, J., Jedich, A., Hoppe, H. U., Harrer, A. (2018). Using Eye-Tracking to Analyze Collaboration in a Virtual Role Play Environment. *International Conference on Collaboration and Technology* (pp. 185-197). Springer, Cham. doi: 10.1007/978-3-319-99504-5_15

Othlinghaus, J., Hoppe, H. U. (2016). Supporting Group Reflection in a Virtual Role-Playing Environment. *International Conference on Intelligent Technologies for Interactive Entertainment* (pp. 292-298). Springer, Cham. doi: 10.1007/978-3-319-49616-0_30

Othlinghaus-Wulhorst, J., Mainz, A., Hoppe, H. U. (2019). Training Customer Complaint Management in a Virtual Role-Playing Game: A User Study. *European Conference on Technology Enhanced Learning* (pp. 436-449). Springer, Cham. doi: 10.1007/978-3-030-29736-7_33

OTHER PUBLICATIONS

The publications listed below describe works that are relevant in the context of this thesis, but they are not included as dedicated chapters.

Emmerich, K., Neuwald, K., Othlinghaus, J., Ziebarth, S., Hoppe, H. U. (2012). Training Conflict management in a collaborative virtual environment. *International Conference on Collaboration and Technology* (pp. 17-32). Berlin, Heidelberg, Springer.

Doberstein, D., Agreiter, N., Bäumer, M, Cui, M., Abdollahzadegan, S., Heidari, D., Jiang, N., Mentzel, M., Zhang, H., Zheng, H., Othlinghaus, J., Hoppe, H. U. (2016). CuCoMaG – Group Reflection Support in Role-Playing Environments. *DeLFI 2016--Die 14. E-Learning Fachtagung Informatik* (pp. 327-329).

Harbarth, L., Delsing, S., Richtscheid, F., Yücepur, V., Feldmann, F., Akhavanfar, M., Manske, S., Othlinghaus, J. & Hoppe, H. U. (2018). Learning by Tagging – Supporting Constructive Learning in Video-Based Environments. In: Krömker, D. & Schroeder, U. (Eds.), *DeLFI 2018 - Die 16. E-Learning Fachtagung Informatik* (pp. 105-116). Bonn, Gesellschaft für Informatik e.V.

CONTENTS

Abstract iii
List of Included Publications v
1 Introduction 1
1.1 Context and Research Objectives 2
1.1.1 Serious Games for the Training of Social Skills 2
1.1.2 Chatbots in Education 7
1.1.3 Research Objectives 12
1.1.4 Methodological Considerations 13
1.2 Synopsis of Included Publications 15
1.2.1 The Framework 15
1.2.1.1 Conceptual Approach 15
1.2.1.2 Technical Approach 22
1.2.2 Case Studies 28
1.2.2.1 Case Study 1: Conflict Management 28
1.2.2.2 Case Study 2: Customer Complaint Management 32
1.2.3 Dimensions of the Design of Serious Role-Playing Games for the Training of Social Skills 38
1.2.4 Contributions of the included publications to different research topics 40
2 A Technical and Conceptual Framework for Serious Role-Playing Games in the Area of Social Skills Training 43
3 Case Study: Conflict Management 65
Using Eye-Tracking to Analyze Collaboration in a Virtual Role Play Environment 65
4 Case Study: Customer Complaint Management (1) 79
Supporting Group Reflection in a Virtual Role-Playing Environment 79
5 Case Study: Customer Complaint Management (2) 87
Training Customer Complaint Management in a Virtual Role-Playing Game: A User Study 87
6 Conclusion 103
6.1 Additional Results 103
6.2 Summary 109
6.3 Future Perspectives 113
6.3.1 Further developments 113
6.3.2 New implementational approaches 113
6.3.3 Theory- and evidence-based serious game design 114
7 References 115

1 INTRODUCTION

These days, there is a broad interest in the potential of serious games as vehicles for learning and training. In recent years, they have been established as a useful and efficient medium in education and professional practice (Michael & Chen, 2006) (Marr, 2010), and can be utilized for a broad spectrum of application areas, e.g., military, government, education, corporate, and healthcare (Susi, Johannesson, & Backlund, 2007) (Marr, 2010) (Marsh, 2011). Social change and advancing digitalization increasingly demand new and modern learning methods, which is why this area of research is particularly interesting and contemporary. Utilizing the engaging and motivational potential of video games may change the way people learn, and it may make learning more enjoyable (Marsh, 2011).

Combining the serious games approach with the element of role play is particularly promising, as virtual role-playing games provide mobile, safe, and repeatable environments for learners in which they can assume roles in particular contexts, explore new situations, and learn how to act and react without fear of consequences in the real world (Martens, et al., 2008). While traditional role plays can be time-consuming, expensive, difficult to administrate, and are lacking repeatability (Totty, 2005), their virtual counterparts are time and place independent, easy to apply, and can be repeated as often as required as no additional resources are needed. Serious role-playing games provide a unique environment for immersive and interactive learning. They use experiential and situated learning to allow learners to utilize existing and construct new knowledge, try out different problem-solving strategies, and gain experience (Clayton, 2017) in the sense of a learner-centered constructivist approach (Lim, et al., 2009). Serious role-playing games are very versatile and cover numerous areas of learning. They provide instant feedback leading to an increased motivation and user engagement, which is a “key predictor of success” in any learning or training software (Dell'Aquila, et al., 2017). Virtual role plays have the potential to be much more effective than conventional approaches in settings where the social component is a crucial factor (Lim, et al., 2009), which makes them an ideal tool for the training of social skills. Improving social skills is important for people’s personal and professional development since these skills help them to become better relationship builders, and the quality of these relationships plays a very important role in an individual’s life, the life as part of a society, and also workplace life (Gökel & Dağlı, 2017), which is why they are the subject of many professional training programs and pose a relevant object of research (Dell'Aquila, et al., 2017).

Shaping the pedagogical outcomes constitutes a special challenge for serious role-playing games. The effects of a learning experience usually depend on feedback and post-role-play reflection; without this, the transfer to real-world settings cannot be ensured (Lim, et al., 2009). While traditional training scenarios typically rely on video recording and note-taking, virtual learning environments can provide structured recordings with integrated indexing, navigation instruments, search functions, and cross references between different media and data sources. Computer-supported analyses can help track and evaluate the learners’ performance (Othlinghaus-Wulhorst & Hoppe, 2020).

This thesis presents a conceptual and technical framework for the design and implementation of serious role-playing games for the training of specific social skills in virtual 2D learning environments involving AI-controlled chatbots in dialog-centric settings. The following Chapter 1.1 provides the context of this research work and outlines research objectives and methodological considerations in detail. Chapter 1.2 summarizes the included publications, explains and substantiates the pillars of the developed framework (Chapter 2), and presents and compares two case studies (Chapter 3-5) developed based on the framework. Chapter 6 presents additional research results, which supplement the findings of the published works, summarizes the main outcome of this thesis, and provides an outlook on future research.

1.1 Context and Research Objectives

This section outlines the goals and research objectives of this thesis, provides a framework for contextualizing the following chapters, and locates them in relevant fields of research. In the first two subsections, serious games for the training of social skills (Section 1.1.1) and the use of chatbots in education (Section 1.1.2) are introduced as the research background for this thesis. In the third subsection, the research objectives are described in detail (Section 1.1.3). The fourth subsection focuses on methodological considerations with regard to the field of serious games evaluation (Section 1.1.4).

1.1.1 Serious Games for the Training of Social Skills

Technology-enhanced learning and teaching has significant advantages, such as the ability to provide new insights and perspectives through visualizations and animations, ubiquitous access to information, as well as opportunities for self-directed and self-regulated learning, exchange, and collaboration (Kickmeier-Rust & Albert, 2010). The serious gaming approach supplements this rich educational potential of learning technologies with the strong motivational potential of computer games. Today, serious games constitute an increasingly important medium for education, training and behavioral or social change (Michael & Chen, 2006). Nowadays, there are many reports in several different domains of using video games to produce changes in people's knowledge, attitude, or behavior that will be transferred to "real life" settings outside the gaming context (Beale, 2011).

Many different definitions of the term *serious games* exist. Most of them agree on a core meaning: it refers to digital games used for purposes other than mere entertainment (Susi, Johannesson, & Backlund, 2007). Zyda defines a serious game as "a mental contest, played, with a computer in accordance with specific rules, that uses entertainment to further government or corporate training, education, health, public policy, and strategic communication strategies" (Zyda, 2005), while Marsh describes serious games as "digital games, simulations, virtual environments and mixed reality/media that provide opportunities to engage in activities through responsive narrative/story, gameplay or encounters to inform, influence, for well-being, and/or experience to convey meaning" (Marsh, 2011). When it comes to the difference between entertainment games and serious games, Zyda argues that it is the addition of *pedagogy* (activities that impart knowledge or skill) that makes games serious (Zyda, 2005).

In summary, several important aspects are linked to the concept of serious games:

Explicit focus on education and training. Serious games are supposed to convey specific knowledge or train certain skills (Susi, Johannesson, & Backlund, 2007). Michael and Chen differentiate between games that educate and games that train (Michael & Chen, 2006). *Games that educate* are games that are intended to convey knowledge or facts processed in a playful way, while *games that train* are supposed to improve specific skills of the learner in virtual environments or simulations (Michael & Chen, 2006). This thesis focuses on the second category.

Fun. The serious gaming approach uses the attractiveness and great motivational potential of digital games for learning and training (Susi, Johannesson, & Backlund, 2007). Using digital games as a "toy" for learning implies that the activity is intrinsically motivating by itself because it is *fun* (Ritterfeld, Cody, & Vorderer, 2009). However, to achieve an optimal learning effect, the fun needs to be linked closely to the learning process (Iten & Petko, 2016). Several studies show that the learners' enjoyment of a serious game has an impact on their motivational gains to continue engaging with the subject matter that is being taught (Iten & Petko, 2016). In a content analysis of 60 game reviews, Wang, Shen and Ritterfeld found five main fun factor categories in serious games: overall game design, visual presentation, audio presentation, complexity and diversity, and control (Wang, Shen, & Ritterfeld, 2009). Complementing these findings, they identified several "super fun factors" that contribute to an especially high level of enjoyment in further trials: narrative-related elements such as character and dialogues, humor, and social interaction (Shen, Wang, & Ritterfeld, 2009).

Positive effects on the learners. Serious games intend to facilitate deep and sustained learning (Gee, 2007). Serious games allow learners to experience situations or working conditions that are hard or even impossible to experience in the real world for safety, cost, or time reasons (Corti, 2006) (Susi, Johannesson, & Backlund, 2007). They do not only have a positive effect on the development of the players but can be conducive to several different skills and competences related to cognitive, social, analytical, and strategic aspects (Mitchell & Savill-Smith, 2005) (Squire & Jenkins, 2003). There is evidence that serious games are more efficient than traditional pedagogy or other educational technologies (Prensky, 2000) (Ritterfeld, Cody, & Vorderer, 2009) (Zhonggen, 2019). Several meta studies indicate that serious games improve learners' academic achievements, encourage participation in learning activities, and increase positive affect toward learning (Zhonggen, 2019) (Lamb, Annetta, Firestone, & Etopio, 2018). However, the effectiveness of a serious game depends on how well it has been designed. Serious games may not be as effective as intended if social psychological theories are not appropriately applied. The educational outcome of a serious game depends on several factors like capturing the attention of the learners and reinforcing the desired behavioral changes by showing that these are advantageous (Susi, Johannesson, & Backlund, 2007). What is also important is that learners are emotionally involved and affected by the content and identify with virtual characters in the game (Susi, Johannesson, & Backlund, 2007). If some of these elements are absent, the education outcome can be limited or non-existent (Susi, Johannesson, & Backlund, 2007).

Learning objectives. The purpose of serious games is to induce changes in the learners' attitudes and behaviors (learning objectives), and therefore game designs are efficacious only to the extent to which they achieve such changes (Beale, 2011). According to Beale (Beale, 2011), these behavioral changes specify the targeted outcome of the game and should be concrete and measurable, and generally include either:

a) improving the learner's performance of a particular skill,
b) increasing the learner's knowledge about a particular issue,
c) altering the learner's attitude towards a particular issue, or
d) altering the perceived experience of a particular event.

In terms of skill training, boundaries in the game are often initially set rather broadly, but as the learner progresses, the requirements are made more specific. The game is shaping the trained skill over time, working towards the final version of what is defined as an acceptable performance (Beale, 2011). Mapping learning objectives to game mechanics is one of the biggest challenges in the design of serious games. Pedagogy and teaching strategies need to be properly aligned to game activity and assessment in order to balance game features with pedagogical aspects and make sure that the learning objectives will be achieved (Lameras, et al., 2017).

Amongst several other fields of application, serious games have extensively and effectively been used in job skills training to sustain the development of work-related competences, knowledge as well as social and personal attributes (Earp, Ott, Popescu, Romero, & Usart, 2014). This thesis focuses on serious games for the training of social skills based on role plays in a workplace-oriented context. It has already been explained what serious games for training are; the following passage will therefore focus on the other two components: *social skills* and *role play*. Social skills can be seen as a sub-set or sub-category of soft skills, while the term *soft skills* characterizes a broad concept describing a set of personal attributes or traits that express how people know and manage themselves and also their relationships with others (Dell'Aquila, et al., 2017). No universal definition of the concept "soft skills" exists, but Dell'Aquila et al. merge several different approaches into the following comprehensive definition: "Soft skills are not domain or practice specific; experientially based; both self and people orientated; goal-related behaviors; inextricably complementary to hard technical knowledge and skills enabling completion of activities and accomplishment of results; and crucial for effective leadership performance." (Dell'Aquila, et al., 2017).

The sub-category *social skills* refers to soft skills that are related to interaction with other people. Combs and Slaby describe it as "the ability to interact with others in a given social context in specific ways that are socially acceptable or valued and at the same time personally beneficial, mutually beneficial, or beneficial primarily to others" (Combs & Slaby, 1977). Riggio describes seven basic social skill dimensions involving skills in sending, receiving, and controlling communication in two domains (emotional-nonverbal and social verbal) (Riggio, 1986). These seven dimensions are:

1. Emotional expressivity
2. Emotional sensitivity
3. Emotional control

4. Social expressivity
5. Social sensitivity
6. Social control
7. Social manipulation

Emotional expressivity is a general skill in nonverbal sending and refers to a person's ability to express emotional states as well as attitudes and cues of interpersonal orientation. Persons with this skill are animated, energetic, and can be characterized as emotionally charged. They may be able to emotionally arouse or inspire others because they are able to transmit their emotional states. *Emotional sensitivity* refers to the ability to receive and decode nonverbal communications of other people. Persons with this skill are vigilant in observing the nonverbal behavior of others and are able to decode it rapidly and efficiently; as a result, they may be more susceptible to becoming emotionally aroused by others. *Emotional control* describes the general skill to control and regulate emotional and nonverbal displays, e.g., being able to display emotions on cue and to use conflicting emotional cues to mask emotional states. People with high emotional control are likely to be good emotional actors and may tend to suppress the display of (even spontaneous and extreme) emotional states. *Social expressivity* refers to a verbal speaking skill allowing to engage others in social interaction. Persons with this skill appear to be outgoing and gregarious, and they are usually good at initiating conversations with others and speaking spontaneously—sometimes without apparent monitoring of content. *Social sensitivity* is the ability to decode and understand verbal communication as well as having a general knowledge of social behavior norms and rules. Socially sensitive persons are attentive to others (i.e., they are good watchers and listeners). They may become overly concerned with the appropriateness of their own behavior as well as the behavior of others due to their knowledge of social norms. The general skill in social self-presentation is called *social control*. Persons with high social control appear to be tactful, socially adept, and self-confident. They are able to play various roles and adjust personal behavior to fit what they consider appropriate in any social situation. *Social manipulation* is not only a social skill—it can also be seen as a general attitude or an orientation. Persons with this ability can manipulate others or alter elements of a situation to influence the outcome of social encounters when they consider it necessary and useful. This does not mean that they only try to achieve outcomes benefiting for themselves; they may even act in a self-sacrificing manner (e.g., taking responsibility to protect someone else).

Concrete important social skills include, for example, conflict management, negotiation, leadership, personal effectiveness, active listening, creative problem-solving, strategic thinking, decision-making, and team building, as well as influencing and selling skills (Dell'Aquila, et al., 2017). Improving social skills is important for both the personal and professional developments of people since these skills help them to become better relationship builders—and the quality of these relationships plays a significant role in an individual's life, the life of a society, and also workplace life (Gökel & Dağlı, 2017). Appropriate social behavior may be even more important than academic or job skills when it comes to determining whether someone is perceived as a competent person (Black & Langone, 1997). Social Skills (sometimes also referred to as "emotional intelligence") in a workplace context may influence the success in interacting with colleagues or principals, the strategies people

use to manage conflict and stress situations, and overall job performance (Brackett, Rivers, & Salovey, 2011). Social skills are essential for a successful career development, particularly in social professions (Otto, et al., 2019).

A growing interest from researchers, managers, representatives of industry, commerce and organizations, and educators can be noticed; they regard soft skills as crucial for fostering personal and collective growth, and the creation of new practices in professional, vocational, and educational contexts (Dell'Aquila, et al., 2017). Usually, attempts at improving social skills are made by applying skill-training practices (Gökel & Dağlı, 2017). These trainings often use a behavioral approach to improve communication, decision making, problem solving, self-management, self-control and competitive skills, and they are usually performed in group settings (Gökel & Dağlı, 2017). One vital and prominent instrument of training social skills is *role play*. Assuming roles allows learners to explore new situations and to train how to act and react in these situations. Vice versa, observing role play can lead to conclusions about the observer's own behavior (Martens, et al., 2008).

Social skills training through role play covers several aspects of Bandura's *social learning theory* (Bandura, 1986), which assumes that the social context is the dominant driver for learning, e.g., by observation and dialog. For example, role play emphasizes the difference between the acquisition and the performance of behavior as well as the importance of providing secure settings for experimenting with new behavioral strategies (Lim, et al., 2009). It enables the creation of knowledge and meaning from concrete experiences (although these experiences are imagined) (Lim, et al., 2009).

According to John Dewey, a famous educational theorist, the learning process is an improvement of knowledge through continuous training ("learning by doing") (Dewey, 1938). This "learning by doing" involves active participation in a planned and controlled event, a reflective analysis on the experience, and the application of learned principles to real-life situations. Dewey postulates that experience is a rich resource for learning that helps learners to understand and retain knowledge in persistent, unforgettable ways (Dewey, 1938). Role play meets these criteria perfectly and enables participants to directly experience a situation, learn the consequences of their actions, and improve by continuous training.

Another firm theoretical basis for learning through role play is Kolb's *experiential learning theory* (Kolb, 1984). Similar to Dewey, Kolb assumes that the learner must be actively involved in the experience of learning. Furthermore, he suggests that learning is cyclical and made up of four stages: concrete experience (having an experience), reflective observation (reflecting on the experience), abstract conceptualization (concluding from the experience) and active experimentation (trying out what has been learned). The learning cycle therefore provides feedback, which is the basis for new actions and evaluation of the consequences of these actions (Healey & Jenkins, 2000). Role-playing allows participants to immerse themselves in a learning environment and undergo all different stages of learning.

However, traditional role play trainings tend to be time-consuming, expensive, difficult to administrate, and lack repeatability (Totty, 2005). In contrast, their virtual counterparts are time and place independent, easy to apply, and can be repeated as often as required as no additional resources are needed. Serious role-playing games provide a unique environment

for immersive and interactive learning. They use experiential and situated learning to allow learners to utilize their existing and construct new knowledge, try out different problem-solving strategies, and gain experience (Clayton, 2017) in the sense of a learner-centered constructivist approach (Lim, et al., 2009).

A comprehensive list of serious role-playing games for social skills training can be found in Chapter 2. In summary, they can be assigned to three main categories of relevant social skills: (1) leadership skills, (2) communication skills, or (3) conflict management. Main distinctive characteristics of the different games are: mode (singleplayer vs. multiplayer), learning objective (teamwork, leadership styles, negotiation/communication skills, conflict resolution, etc.), and underlying framework/model/theory (educational programs and curricula, psychological theories, etc.) (Othlinghaus-Wulhorst & Hoppe, 2020). It can be concluded that a large number of serious role-playing games is available on the market, but many of them are not built on psychological theories and models and are thus lacking a solid theoretical background and conception. Furthermore, there are several existing models and frameworks for the general design of serious games describing fundamental components of such systems with the intention to support formal approaches to game design, but these models are defined on a very general level and lack concrete design or evaluation guidelines (Othlinghaus-Wulhorst & Hoppe, 2020). Thus, the framework developed within the scope of this thesis (presented in Chapter 2) is trying to fill this gap and provide a concrete scaffold for the design and implementation of serious role-playing games for the training of social skills in dialog-centric settings using chatbots in order to support more efficient and effective design and implementation of such systems.

1.1.2 Chatbots in Education

Chatbots are messaging-based intelligent conversational agents that communicate with users by using natural language. The purpose of chatbots is to simulate a human conversation via text or voice interaction. Today, chatbots are used for a number of purposes, such as customer service, social and emotional support, information, e-commerce, entertainment, healthcare, and more (Kerly, Hall, & Bull, 2006) (Shawar & Atwell, 2007) (Brandtzaeg & Følstad, 2017). Recent advances in natural language processing and machine learning, as well as the creation of social media platforms like Facebook and Telegram, and intelligent personal assistants like Apple's Siri or Amazon's Alexa triggered a new hype around chatbots, which lead to rapid developments within industry and research in this field (Braun & Matthes, 2019). Major companies like Google, Facebook, and Microsoft see chatbots as *the* next popular technology. In 2016, Facebook and Microsoft provided resources for the creation of chatbots to be integrated into their respective messaging platforms. The following year, more than 30,000 chatbots were launched on Facebook Messenger (Brandtzaeg & Følstad, 2017). In January 2019, the number of active chatbots increased to more than 300,000 (Zhang, Oh, Lange, Yu, & Fukuoka, 2020), many of them for education and learning. Other messaging platforms have seen a significant increase in chatbots as well.

One of the earliest chatbots was developed in 1966 by Joseph Weizenbaum at the MIT Artificial Intelligence Laboratory (Weizenbaum, 1966). It was named ELIZA and was based on a very simple pattern matching and substitution methodology to simulate understanding of

the human user and engage in a conversation. It relied on scripts originally written in MAD-Slip to process user inputs and generate appropriate responses (Weizenbaum, 1966). From user input, certain keywords are extracted and compared with a stored list of (syntactic) patterns for the respective keyword. In case of matches, a sentence referring to the keyword is returned or otherwise a fallback response is generated (Storp, 2002). The search for keywords is to be understood as a simple character matching. Words are strings without meaning for ELIZA, and they can occur in certain syntactic contexts (Storp, 2002). The most famous script (called the "DOCTOR" script) included a simulation of a psychotherapist engaging in a conversation with a human patient. Weizenbaum chose this role because he considered it easy to implement: The therapist's activity is essentially based on paraphrasing the patient's statements as a question, "mirroring" them, thereby motivating the patient to continue talking (Storp, 2002). The therapist completely abstains from own conversational initiatives and reacts only to what is said.

While Weizenbaum saw ELIZA primarily as proof of the context-dependence (in the sense of syntactic context) of language understanding (Weizenbaum, 1976), the program was regarded as a breakthrough in machine understanding of natural language. There has been a lot of progress since the development of ELIZA and new methods of natural language processing have been developed in the past decades, but it still exists today in numerous versions and can compete with some of the modern chatbots even after several decades. Thus, many modern chatbots can be considered direct evolutions of ELIZA (Storp, 2002). From today's perspective, the capabilities of this system seem rather modest, but ELIZA marked the beginning of the experimental phase of machine processing of natural language and still serves as inspiration for many modern chatbots (Shawar & Atwell, 2007). After Weizenbaum's ELIZA, the first programs emerged (in the field of AI) with which increasingly sophisticated language processing tasks could be performed (Menzel, 2003). One of these is the *Artificial Linguistic Internet Computer Entity* (A.L.I.C.E.), which has been originally composed by Richard Wallace and the Alicebot free software community between 1995 and 2000 (Wallace, 2004). It is based on the *Artificial Intelligence Markup Language* (AIML), an XML language Wallace specifically designed for creating stimulus-response chatbots (Wallace, 2009). A.L.I.C.E. is considered one of the strongest programs of its type and won the *Loebner Prize*, an annual Turing Test, in 2000, 2001, and 2004 (Bradeško & Mladenić, 2012). Data objects in AIML consist of units called topics and categories. While *topics* represent optional top-level elements that serve as context, *categories* are the basic knowledge units in AIML. They consist of patterns (input question) and templates (output answers), and can be related to one of the specified topics (Shawar & Atwell, 2005). The *patterns* are matched against the user input and *templates* are used to generate the chatbot response. Besides these basic elements, AIML offers several additional features like wildcards, recursion, and different control structures, which allow a fairly complex and sophisticated chatbot design.

Many modern chatbots are still built on such modular stimulus-response systems like ELIZA and A.L.I.C.E that match linguistic input with an internal pattern database to generate appropriate responses. This category of chatbots is called *retrieval-based* (Winkler & Söllner, 2018). The centerpiece of these chatbots is still the so-called *knowledge database*, the part of the system in which recognition patterns, keywords, and answers are stored (Storp, 2002).

The actual program controls the conversation flow, i.e., coordinates input and output, the activation of the knowledge database and, if necessary, other modules such as the output of spoken language. However, the individual systems differ in the flexibility and size of their knowledge bases and in the performance of their control programs: While some chatbots communicate only in written language, some can generate spoken responses and are able to understand spoken language input as well. Depending on their input and output channels, chatbots can be categorized into speech-based, text-messaging, and multimodal chatbots (Jain, Kumar, Kota, & Patel, 2018). There are systems that, like ELIZA already, generate responses from stubs. Others, however, randomly select from a set of pre-formulated responses. Some chatbots compare only a part of an input with the sample database and may disregard the rest. Others analyze all parts of the input and combine the answers to the respective single parts to one output. Chatbots also differ in their ability to "memorize", i.e., whether parts of input can be stored in variables and recalled in the course of the conversation. Furthermore, there is the question of learning ability. Most chatbots (e.g., A.L.I.C.E) keep a protocol of their conversations and suggest new recognition patterns or keywords based on them. A human decides whether these are then added to the knowledge base. This process is called *supervised learning*. Modern systems can even expand their database independently based on the protocols. Some bots can also use external data sources if there is no recognition pattern for the input (Storp, 2002).

New chatbot technologies rely on advanced techniques, such as information extraction and machine learning (Hristidis, 2018). The so-called *generative* chatbots (in contrast to retrieval-based) do not use pre-defined responses; instead, they generate responses from the input by using machine learning techniques (Winkler & Söllner, 2018). Generative models have the potential to generate proper responses on the fly, which potentially make the chatbot appear more human and capable of longer dialogs (Winkler & Söllner, 2018). On the other hand, generative-based chatbots are difficult to implement and operate in comparison to retrieval-based chatbots (Hussain, Sianaki, & Ababneh, 2019) and need to be trained on large datasets of question-answer pairs (Kapočiūtė-Dzikienė, 2020).

All these approaches to natural language understanding have to face some critical issues that are grounded in the nature of human language (Storp, 2002). The faculty of speech is an essential component of human intelligence. Even if dissociations between language mastery and general intelligence can be observed, language and intelligence nevertheless develop in close interaction (Menzel, 2003). For AI research, this raises the question of whether machines can think like humans or at least talk like humans and how this ability can be transferred to machines. In order to make computers talk, it must first be clarified how language is structured and how it works. It seems necessary to trace natural language back to theoretical and mathematical models, which can then be made accessible to the computer (Storp, 2002). Natural language is characterized by various forms of ambiguity. In contrast, artificial language systems, such as programming languages, are characterized by their immutability and uncompromising unambiguity. In comparison with such artificial languages, the characteristics of natural language stand out clearly (Storp, 2002):

- Natural language offers an infinite number of possible combinations on the basis of a limited set of elements. Morphemes are combined to form new words that are

understood even if they are not yet lexicalized; the possibilities for sentence formation are immense despite syntactic limitations.

- Natural language distinguishes a variety of speech acts. This results in pragmatic ambiguity: thus, the sentence “Is the door open?” may be a simple question, an indirect request or a reproach.
- Natural language utterances follow linguistic economy principles and are therefore often underspecified. The process of comprehension thus involves the reconstruction of implicit information content.
- Natural language expressions can be used in a non-proper sense: Metaphor and metonymy, irony, etc.

Besides these difficulties, the main problem of machine processing is the ambiguity of natural language, which can be (Storp, 2002):

- Lexical ambiguity (polysemy and homonymy)
- Structural/syntactic ambiguity
- Referential ambiguity (pronouns cannot be assigned on the basis of linguistic context alone)
- Pragmatic ambiguity

All these characteristics of natural language pose problems and need to be considered when implementing chatbots.

While chatbots were initially developed primarily for entertainment purposes, today there are also many other fields of application, such as education, search engines, commercial applications, and e-commerce (Shawar & Atwell, 2007). Past studies show that they can be successfully used for learning and present feasible means to improve learners’ results (Kerly, Hall, & Bull, 2006) (Winkler & Söllner, 2018). Chatbots have a long history as pedagogical agents used in educational settings (Smutny & Schreiberova, 2020). From the early 1970s, chatbots within digital learning environments known as *intelligent tutoring systems* (ITS) have been developed (Laurillard, 2013). This type of system is characterized by giving learners some kind of natural language user interface that allows them to enter the steps required for solving a specific problem (VanLehn, 2011). The intelligent tutor can give feedback and hints on each of the learners’ steps. The goal of ITS is to engage the learners in sustained reasoning activity and to interact with them based on a deep understanding of their behavior (Corbett, Kenneth, Koedinger, & Anderson, 1997). Over the last decade, chatbots have been more and more incorporated into the educational area, which implies an increased interest in the ways chatbots might be used for teaching and learning. Useful chatbot systems can provide a range of benefits for learners based on their ability to respond in an instant and natural manner, particularly when the interaction is embedded within a broader context of an integrated learning environment (Smutny & Schreiberova, 2020). Application fields for chatbots in the context of education and learning are numerous, e.g., language learning, mathematics, history, economics, politics, memory and logic training, medical education and therapy, and more (Winkler & Söllner, 2018) (Smutny & Schreiberova, 2020). AI-controlled chatbots are becoming an increasingly popular tool for the training of skills requiring a self-paced approach and sustained practice (Schussler, Frank, Lee, & Mahfouz, 2017). They have also been used in

the context of serious role-playing games in the area of social skills training, for example to train job interview skills (Malzahn, Buhmes, Ziebarth, & Hoppe, 2010) (Hoque, Courgeon, Martin, Mutlu, & Picard, 2013), leadership skills (Knode & Knode, 2011), communication skills (Hubal, Frank, & Guinn, 2003) (Lane & Hays, 2008) (Flynn, McKinnon, Bacon, & Webb, 2011) (Vaassen & Wauters, 2012) (Linssen, de Groot, Theune, & Bruijnes, 2014) (Augello, Gentile, & Dignum, 2016), and conflict management (Mateas & Stern, 2003) (Glock, et al., 2011) (Linssen, de Groot, Theune, & Bruijnes, 2014).

The use of chatbots in serious role-playing games has several advantages:

- Having the learner interact with a chatbot instead of a human actor ensures a certain level of standardization that could never be achieved in a setting that relies on humans (Othlinghaus-Wulhorst & Hoppe, 2020).
- Scenarios with chatbots are repeatable, time and place independent, and no additional resources are required (Othlinghaus-Wulhorst & Hoppe, 2020).
- It is possible to simulate critical situations (e.g., bullying incidents, violence, or medical emergencies) without causing harm to actual people in the process (Schussler, Frank, Lee, & Mahfouz, 2017).
- Virtual role-playing in general allows learners to test boundaries more than in a real live social setting. The “magic circle” of game with computer-controlled characters may be more clearly defined than with human actors in a role-playing scenario (Spierling, 2008).
- Through text-only interaction, it is possible to explicitly judge phrasing, as real world parameters are missing that might have disguising effect on recognition (Spierling, 2008).

On the other hand, there are some problems and pitfalls connected to the use of chatbots in serious role-playing games:

- A chatbot can only be as good as the knowledge base it is built on to generate answers (Abdul-Kader & Woods, 2015). A problem of “classic” chatbots is the fact that they do not keep track of the conversation and have no real understanding of the answers. However, chatbots need to show a realistic and responsive behavior in order to increase the learners’ engagement and to be conducive to the immersive nature of role plays (Othlinghaus-Wulhorst & Hoppe, 2020).
- Purely text-based interactions are missing non-verbal cues like body language and presence. Several emotional levels are omitted, and the stress a learner is experiencing in a “real” situation may also be missing (Spierling, 2008).
- The quality of the chatbot implementation is crucial to a successful learning experience, thus it needs to be built on expert knowledge of the respective domain and chatbot development. Otherwise, the designed scenario may result in learning mistakes (Spierling, 2008).

Apart from technical and context-dependent premises, Winkler and Söllner found that individual differences of learners affect chatbot-supported learning and influence the learning outcome (Winkler & Söllner, 2018). These include the attitude and trust towards technology

(i.e., learners with a positive attitude towards chatbots feel more content in chatbot-supported learning), learning characteristics such as trait emotions and personality traits (i.e., trait emotions such as anger, anxiety, or joy and stable personality traits such as agreeableness, conscientiousness, and neuroticism have significant influence on chatbot-directed emotions), educational background and social and technological skills (i.e., learners with better technological skills and a stronger need for interaction benefit more from chatbots) as well as self-efficacy and self-regulated skills (i.e., learners with higher self-efficacy and well-developed self-regulated skills achieve better process quality and learning outcomes).

In summary, chatbots constitute a vital, expedient, and increasingly popular instrument that can be (and has been) successfully used in context of education and learning – especially within serious role-playing games – as it presents feasible means to support learning and improve learners' results. The use of chatbots has several advantages over real enactment, e.g., in terms of standardization and repeatability. However, there are some problems and possible pitfalls that need to be considered when creating chatbots used in the context of a serious game in order to support immersion and avoid learning mistakes. Particularly the quality of the chatbot implementation is crucial to a successful learning experience.

1.1.3 Research Objectives

The main research objectives of this thesis are:

1. **Development of a technical and conceptual framework**

The development of a technical and conceptual framework for serious pole-playing games in the area of social skills training constitutes a major pillar of this work. The framework is specifically designed for the training of particular (mainly work-related) social skills (depending on the concrete use case and scenario) in virtual 2D learning environments involving chatbots in dialog-centric settings. There are several existing frameworks and models for the general design of serious games, but these represent only very general high-level approaches and describe fundamental components of serious games. However, none of them include concrete design or implementation guidelines and they are not specifically designed for the training of social skills[1]. The aim of this work is to provide a comprehensive conceptual and technical framework for the concrete design and implementation of serious role-playing games for the training of social skills in dialog-centric settings with virtual characters, supporting more efficient and effective design and implementation of such game environments. A task that is connected to this objective is the identification of important design dimensions and major challenges related to the design and development of serious games in this field. Furthermore, it is analyzed how reflection processes can be supported in such environments.

2. **Development and evaluation of case studies**

Another major objective of this thesis is to present (both single- and multi-user) case studies based on the proposed framework and to evaluate them with regard to the following research questions:

[1] A list of available frameworks for the design of serious games can be found in Chapter 2.

- Are the developed scenarios perceived as realistic and can they be used in real training situations?
- Do the created chatbots behave as desired and how are they perceived by the learners?
- How can player performance and collaboration quality (in a collaborative setting) be assessed?
- Is there a correspondence between the convergence of visual foci of attention and collaboration quality (in a collaborative setting)?

An overarching objective of this thesis is to provide information to facilitate the design and evaluation of serious role-playing games. In general, the evaluation studies pursue the goal to gain insights regarding playability, usability, and perception of the prototypes following a mixed-method approach, while each of the evaluation studies focuses on different key aspects and specific research questions depending on the particular scenario, use case, and modalities.

1.1.4 Methodological Considerations

Finding out whether a serious game is efficacious in reaching its learning goals is called *outcome evaluation*, while finding out whether particular components of game design have the planned effects on the goal is called *process evaluation* (Beale, 2011). Both should be an essential part of the game development and evaluation process. One of the special features of serious games is that they provide excellent environments for mixed-method data gathering (triangulation) (Mayer, et al., 2014). Mixing qualitative and quantitative methods allows a more complete and synergetic utilization and evaluation of data than separate quantitative and qualitative data collection and analysis do (Wisdom & Creswell, 2013). It is a dynamic option to expand the scope and improve the analytic power of studies (Sandelowski, 2000). The combination of subjective and objective measures is particularly important in the evaluation of serious games as on the one hand, game-based learning is closely related to learners' feelings and (intrinsic and extrinsic) motivation, but on the other hand, objective data is required especially when criteria such as "performance" need to be considered (Wiemeyer, Kickmeier-Rust, & Steiner, 2016).

For the two serious games developed in the scope of this thesis, evaluation data were gathered through mixed-methods, mostly combining pre- and postgame questionnaires and interviews among the users (subjective measures), as well as protocols or transcripts of the conversations and game results (objective measures). In the collaborative scenario, eye tracking has been used as an additional data source. This combination of subjective and objective measures constitutes a comprehensive methodological approach and allows conclusions that could not be reached by a single approach.

Subjective measures. The main questionnaires used in the evaluation studies include the game experience questionnaire (GEQ) developed by Ijsselsteijn et al. (Ijsselsteijn, De Kort, & Poels, 2013) and a questionnaire concerning the human-like qualities of the chatbots developed by Holtgraves et al. (Holtgraves, Ross, Weywadt, & Lin, 2007). The GEQ is a standard instrument for evaluating gameplay experiences. It comprises 42 items across 7 dimensions (challenge, competence, flow, immersion, tension, positive affect, and negative affect). The

seven dimensions can be described as follows (Johnson, Gardner, & Perry, 2018): Sensory and imaginative *immersion* reflects aspects of how strongly players feel connected with the game. This dimension describes the state of reduced self-awareness of the player. It occurs when the software environment is captivating and interesting for the player. *Tension* refers to specific feelings of anxiety, irritation, and pressure related to the game. *Competence* relates to how well players judge their own performance against the goals of the game. Participants can indicate whether they felt skillful and successful during the game. *Flow* indicates whether players lost track of their own effort and the passage of time during the game. *Positive affect* is related to positive emotional experiences (fun, enjoyment, relaxation) and *negative affect* is related to negative emotional experiences (boredom, frustration, disappointment) during the game session. *Challenge* indicates the degree to which players find a game to be difficult or challenging. Ratings are applied on a 5-point Likert scale: "0-not at all", "1 -slightly", "2 -moderately", "3 -fairly", and "4 -extremely". The Holtgraves questionnaire for the rating of human-like qualities of chatbots includes 7 dimensions consisting of bipolar adjective pairs, while each pair needs to be rated on a 9-point scale with the positive adjective corresponding to the value 9 and the negative to the value 1. On these scales, participants can indicate how *comfortable* they felt in a chat situation (from 1 = extremely uncomfortable to 9 = extremely comfortable), as well as the extent to which they perceive their chat partner as *human* (from 1 = extremely nonhuman to 9 = extremely human), *skilled* (from 1 = extremely unskilled to 9 = extremely skilled), *thoughtful* (from 1 = extremely unthoughtful to 9 = extremely thoughtful), polite (from 1 = extremely impolite to 9 = extremely polite), responsive (from 1 = extremely unresponsive to 9 = extremely responsive), and engaging (from 1 = extremely unengaging to 9 = extremely engaging). Apart from those two questionnaires, additional ones related to the specific scenario and research questions have been used, e.g., usability tests (Brooke, 1996), the EduTechRPG questionnaire (Dell'Aquila, et al., 2017), a questionnaire for the measurement of interpersonal attraction (McCroskey & McCain, 1974) and group awareness (Mock, 2017), as well as several self-developed questionnaires. Another subjective measure used within the scope of this thesis is the qualitative interview. Interviews are among the most familiar strategies for the collection of qualitative data (DiCicco-Bloom & Crabtree, 2006). The interviews were pre-structured. The main purpose of the interviews was to gain detailed insights into the game experience of the participants, i.e., to a) understand how they perceive the conversation with the chatbot and their own performance as well the performance of the chatbot and b) identify possible system-related problems and potential improvements.

Objective measures. One important objective measure is player performance or success in the game. Player performance is a complex concept comprising results and processes of action and interaction in and with the game (Wiemeyer, Kickmeier-Rust, & Steiner, 2016). The assessment of player performance is required for several purposes, e.g., to compare the performance of players with and without prior experience in the area of interest or to relate it to certain experimental conditions. For each of the evaluated games, player performance has been operationalized based on several different parameters (e.g., system-generated performance analysis and feedback, total and relative score, completion time, achieved milestones) and involved both data tracked and/or generated by the system and human evaluation of the logged dialog scripts based on clear operational classification schemes.

These clear operational classification schemes are important to ensure objectivity and a high inter-rater reliability.

In summary, serious role-playing games provide excellent environments for mixed-method data gathering and in the conducted evaluation studies make use of mix-methods by combining quantitative and qualitative data as well as subjective and objective measures, which allows a comprehensive and synergetic utilization and evaluation of the collected data.

1.2 Synopsis of Included Publications

This section outlines the main contributions of the publications constituting the following chapters. These chapters can be contextualized according to the investigated research objectives and scopes as outlined in Chapter 1.1.3, as well as the methodological considerations discussed in Chapter 1.1.4.

1.2.1 The Framework

As described in Chapter 1.1.1, there is a broad interest in the potential of serious games as vehicles for learning and training. The training of social skills is one domain of interest and nowadays, there are many reports of the use of serious games in this field of application intended to cause changes in learners' knowledge, attitude, or behaviors that will transfer to "real world" settings outside the context of the game. However, it can be observed that there is a general lack of evidence-based and theoretically driven models for creating and evaluating serious games (in general, but also in this specific field), which poses a major problem for the development of serious games. Without a solid theoretical foundation and the use of well-proven learning strategies, design principles and development methods, a) the transfer of acquired knowledge and skills to the real world cannot be ensured and b) it is impossible to assess the efficacy of any serious game.

There is a number of existing models and frameworks for the general design of serious games that describe their fundamental components and support formal approaches to the design, but as shown in Chapter 1.1.1, most of them are defined only on a very general level and without concrete design or implementation guidelines[2]. The framework presented in Chapter 2 (Othlinghaus-Wulhorst & Hoppe, 2020) is a comprehensive and concrete conceptual and technical framework for the design and implementation of serious role-playing games for the training of social skills in dialog-centric settings with virtual characters.

1.2.1.1 *Conceptual Approach*

On the conceptual level, chat-like interaction with AI-controlled chatbots, the relation between immersion and reflection, as well as adaptive feedback characterize the framework. These three components will be illustrated in the following paragraphs:

Chat-like interaction with AI-controlled chatbots. The benefits of using chatbots in serious role-playing games have been outlined in Chapter 1.1.2. Past studies show that chatbots can be successfully used for learning and present feasible means to improve learners' results

[2] A comprehensive summary of existing models is provided in Chapter 2

(Kerly, Hall, & Bull, 2006). The framework focuses on text-based, chat-like interaction with chatbots as this form of interaction proved to be advantageous. A study by Krämer, Bente, Eschenburg, and Troitzsch (2009) showed that compared to a speech-based or an embodied conversational agents (ECA)-based interface, the text interface was rated as more efficient and easier to use (Krämer, Bente, Eschenburg, & Troitzsch, 2009). The subjects felt less uncomfortable with the exclusively text-based interaction, and the purely text-based interaction has been perceived as more efficient and usable. The chatbots developed within the scope of this thesis are based on AIML. Although AIML has limited capabilities, it presents a functional and efficient means for creating chatbots in educational contexts. AIML is easy to use, configure, and deploy, and does not require large sets of training data. However, the quality of an AIML chatbot depends 100% on the elaborateness of its knowledge base, the AIML scripts. Furthermore, the purely passive nature, the limited expressive capabilities, and the simple pattern-matching mechanism of AIML require workarounds and a restriction on input. The need for such a restriction was, amongst others, identified in the course of the work of Mori et al. (Mori, Berta, de Gloria, Fiore, & Magnani, 2013). There is a risk that a high level of frustration can occur for players when given the opportunity for completely free input in a chat-based serious game, as it is not feasible to cover all the different input possibilities. With so-called *sentence openers* as an aid, this problem is avoided. Sentence openers are predefined sentence fragments for the player to use and complete with text input. The sentence openers represent a kind of scaffolding mechanism on the one hand and simplify the interpretation of the player input on the other hand. Through the sentence openers, the conversation can be structured without strictly defining dialog sequences like in graph-based conversations. In addition, they allow to define the general gist of a message (e.g., affirmation, rejection, asking a question), which reduces the number of possible inputs and facilitates the interpretation of the message. Besides the use of sentence openers, there are some more general tricks for the creation of believable and authentic chatbots (Storp, 2002):

- *Response time*: Most chatbots work very fast due to powerful computers. However, since they simulate a human who has to type his input on a keyboard, the output is slowed down.
- *Tolerance of errors*: It can be assumed that people make mistakes when typing messages. Such errors should be anticipated and intercepted to a certain extent.
- *Topics of conversation*: Successful chatbots should be able to bring in topics of conversation and tell little stories in order to direct the conversation into areas in which they have recognition patterns available, as it is impossible to prepare the chatbot for every conceivable user input.
- *Non sequitur*: Even the chatbot with the largest knowledge database will not succeed in finding a suitable recognition pattern for every possible input. The decisive factor for the performance of a chatbot is how it deals with such input.
- *Repetitions*: If a chatbot always responds to the same input with a single standard answer, it does not feel very natural. Thus, it makes sense to have a whole range of answers available for a recognition pattern from which the system selects randomly. A good chatbot will also try to steer the conversation in a different direction.

- *Personality*: It seems promising to give the chatbot a personality, a virtual identity, in order to make it appear more human. The way people communicate with each other, their choice of words, the coding of moods, etc. say a lot about personality.

Furthermore, Jain et al. developed some general guidelines for the design of chatbots (Jain, Kumar, Kota, & Patel, 2018), such as:

- *Clarify capabilities at the start and on-demand*: A chat-like interface is powerful and allows un-restricted or only partly restricted interaction patterns with natural language. This increases the players' expectations for the bot's capabilities. Insufficient visibility of the capabilities and limits of the chatbots might result in dissatisfaction and frustration. To reduce a possible expectation gap, the chatbot should clearly specify what it can do (as part of the introduction and also later in the conversation).
- *Evaluate application-interface match*: Chatbot designers must identify if the application is suitable for a messaging interface, as this might not be the case for all possible scenarios. Conversational or turn-based features should be essential for the application, and it should be restricted to the chat interface (i.e., avoid links to external web pages).
- *Enable dialog efficiency through context resolution*: Designers should aim to improve dialog efficiency by resolving and maintaining context from previous messages, as humans need context dependence and expect connectedness across the whole conversation. To resolve context, a chatbot should proactively ask intelligent questions to reduce the search space and engage the player in a meaningful conversation. Such context resolution skills will be interpreted as properties of an authentic and intelligent chatbot.
- *Consistent personality with small-talk and humor*: Players relate better with a chatbot that exhibits a consistent personality. They expect human-like conversational etiquette from computer-controlled chatbots. It is recommended that chatbot designers enrich a conversation with humor and a large diversity in chatbot responses.
- *Design for dialog failures*: Interaction via a (semi-)free-form messaging interface can cause conversational flows that are not modeled and thus produce dialog failures. Designers should explicitly design for such situations, e.g., by admitting failure and showing a list of capabilities with examples or providing a witty conversational cover-up.

They also developed design implications for chatbot platform design (Jain, Kumar, Kota, & Patel, 2018):

- *Combine text-based interface with buttons and media*: Players find the combination of text with buttons and media contents such as images and videos natural and engaging. It should be possible to open any external content or task in-line. A messaging platform should highlight additional features like menus and buttons.
- *Enable efficient input from users*: A messaging interface should help to reduce interaction cost, e.g., trough auto-suggestion buttons.
- *Provide persistent view on chatbot capabilities and context*: A chatbot platform should enable an effective way to discover the chatbots' capabilities and context.

These guidelines and tricks should be considered when designing a chatbot to be used in a training scenario for learning social skills independent from the chatbot technology and platform used, and they have been used for the chatbot design within the scope of this thesis.

Immersion and reflection. As the educational impact of serious role-playing games strongly draws on the so-called "willing suspension of disbelief" (Coleridge, 1984) of the players committing to the role they are supposed to play, these games need to create a certain degree of immersion (Lim, et al., 2009). According to Nacke, "[…] immersion in the game world derives from the player becoming the game character, in the sense of the player having the experience of acting within the game world" (Nacke, 2009). Murray describes immersion as "A stirring narrative in any medium can be experienced as a virtual reality because our brains are programmed to tune into stories with an intensity that can obliterate the world around us… The experience of being transported to an elaborately simulated place is pleasurable in itself, regardless of the fantasy content. Immersion is a metaphorical term derived from the physical experience of being submerged in water. We seek the same feeling from a psychologically immersive experience that we do from a plunge in the ocean or swimming pool: the sensation of being surrounded by a completely other reality, as different as water is from air, that takes over all of our attention, our whole perceptual apparatus." (Murray, 2017). Being immersed in serious games means that players have "a heightened sense of presence through individual identity, are engaged in the content, and thus are intrinsically motivated to succeed in the challenge of the game's goal" (Annetta, 2010). *Presence* can be understood as the psychological perception of being in the game environment in which the player is immersed (Witmer & Singer, 1998). Players feel like individuals in the game; they have a true identity and feel like they are present in the virtual world of the game (Annetta, 2010). If players reach this sense of identity and immersion, they become motivated to proceed and succeed in the game. This intrinsic way of motivating learners is one of the main benefits of serious games and is certainly something conventional instruction modes do not provide (Yee, 2006) (Kickmeier-Rust & Albert, 2010). Players get immersed in games because they find them intrinsically satisfying and thus become more engaged in the learning task (Annetta, 2010). When players are present, engaged, and motivated to master the game and continue its challenges, they reach a state of flow (Annetta, 2010). Flow is a concept that was first described by Csikszentmihalyi, who defines it as a "state in which people are so involved in an activity that nothing else seems to matter; the experience is so enjoyable that people will do it even at great cost, for the sheer sake of doing it" (Csikszentmihalyi, 1990). He identified eight characteristics of this flow state:

1. *We confront tasks we have a chance of completing;*
2. *We must be able to concentrate on what we are doing;*
3. *The task has clear goals;*
4. *The task provides immediate feedback;*
5. *One acts with deep, but effortless involvement that removes from awareness the worries and frustrations of everyday life;*
6. *One exercises a sense of control over their actions;*
7. *Concern for the self disappears, yet, paradoxically the sense of self emerges stronger after the flow experience is over; and*

8. *The sense of duration of time is altered.*

These characteristics should be seen as guidelines for designing activities, tasks, and scaffolds within serious games (Annetta, 2010).

A study by Cheng, She, and Annetta showed that immersion has a positive impact on learning outcomes – particularly when the player performance is high. (Cheng, She, & Annetta, 2015). Kickmeier-Rust and Albert argue that the aspect of immersion has been rarely captured in many serious games (Kickmeier-Rust & Albert, 2010). The immersive potential of serious games is a key advantage of serious game, but fragile. Thus, it is necessary to provide a subtle balance between challenge and ability. According to Kickmeier-Rust and Albert, immersion is likely to occur when learners experience a balance between the challenges of the game, the learners' ability to master the game, and their current knowledge about it (Kickmeier-Rust & Albert, 2010). Serious games need to provide feedback about the learners' performance and tailor challenges to their current capabilities in order to facilitate immersion (Kickmeier-Rust & Albert, 2010). According to Kiili, games should provide clear goals and appropriate feedback, while inappropriate challenges and bad usability reduce the possibility of immersion and flow experience (Kiili, 2005). It is important that pedagogical and didactic measures do not compromise game experience, immersion, and flow; they should intertwine (Kickmeier-Rust & Albert, 2010).

Based on the findings described above, it is perfectly reasonable to carry out the enactment in an immersive situation—especially in terms of experience-based, authentic learning. However, this does not necessarily apply to the debriefing or reflection phase, which is indispensable for the learning process. Reflection is needed to ensure the transfer of the learning content to real-life situations (Lim, et al., 2009). It is a successful tool to improve learning processes (Jonassen, Mayes, & McAleese, 1993), which enables people to recapture, rethink, and evaluate their experiences in order to develop new understandings and appreciations (Boud, Keogh, & Walker, 1985). Self-reflection is an intentional process in which individuals reflect on and address their ideas and actions related to their real and ideal self-concept (Greif, 2008). In addition, outcome-based self-reflection describes the process in which the persons not only deliberately reflect on themselves, but also develop implications for future actions or self-reflections in the process (Greif, 2008). Conscious self-reflection requires the activation of intuitive self-awareness (Greif, 2008). This means, in order to be able to reflect on their behavior the learners need to step out of their role and adopt a different perspective at this point. In conclusion, there is reason to assume that immersion tends to impede critical self-reflection (Malzahn, Buhmes, Ziebarth, & Hoppe, 2010). It can be assumed that the amount of reactive attention that is required for immersion constricts the learners' ability to distance themselves from their role, which in turn interferes with self-reflection. In order to support role distance and the change of perspective needed for reflection, the presented framework proposes a strict separation of phases of immersion and phases of reflection, which means the actual role-playing game should be separated from the (post-role-play) reflection or debriefing session. This separation is supposed to support the learning process as it is advancing meta-cognitive abilities (Malzahn, Buhmes, Ziebarth, & Hoppe, 2010).

Role distance can, for example, be reinforced by a change of the graphical interface. Instead of showing the scene from the learners' point of view (first-person perspective), it could now be shown from a third-person perspective or bird's eye view, enabling the learners to step out of their role. It can also be helpful to allow the learners to navigate between single conversational phases of the game. In online courses, Verpoorten, Westera, and Specht identified reflection identifiers, several of which can be also applied to other technologically enhanced learning activities like serious games, such as graphical representation of contents, visualization tools, records of marks and remarks, shared annotations, self-assessment (Verpoorten, Westera, & Specht, 2011). As of now, an explicit post-role-play reflection or debriefing phase cannot be found in many serious games, but as explained above, there is strong evidence that it is important for facilitating learning transfer. The debriefing could be performed either through system-generated reflection support or with the help of a human facilitator who discusses the learners' results and performance with them. Both variants have advantages and disadvantages (De Troyer, 2017): While a debriefing moderated by a human expert may be quite effective, it is costly, time consuming, and all involved participants need to be in the same place. On the other hand, there is no general approach for creating automated debriefing or reflection systems for serious games, and the development of such can be a quite complex task, as many different types of serious games exist, which may require different approaches (De Troyer, 2017).

Adaptive Feedback. As stated above, reflection phases are an important part of the learning process, and they have to be adequately supported for successful learning. Feedback on the performance of the learner is necessary to ensure the transfer to real-world settings (Lim, et al., 2009) and to help them improve their performance through receiving information about the correctness of their actions (Shute, 2008). Feedback in games is a crucial thing that is even more important in serious games (Annetta, 2010). Feedback lets learners know where they are in the game's narrative and what they need to do succeed in future challenges within this narrative (Annetta, 2010).

In addition to the influence feedback has on achievement, it is also depicted as a significant factor for motivating learning (Shute, 2008). There is a large body of research on feedback and its features, functions, interactions, forms, and mechanisms.

Johnson et al. identified four characteristics of feedback (Johnson, Bailey, & Van Buskirk, 2017):

1. The *type* of feedback (e.g., outcome-based or process-based feedback)
2. The *timing* of feedback after an action (i.e., immediate or delayed feedback)
3. The *modality* in which the feedback is presented (e.g., spoken or text-based feedback)
4. *Adaptation* to the learner characteristics (e.g., in regards to prior knowledge or spatial ability)

Shute conducted an extensive review of the large corpus of research on feedback and concludes with a set of important guidelines for generating formative feedback to enhance learning, which should be considered when designing learning environments (Shute, 2008). These guidelines include:

Things to do:

- *Focus feedback on the task, not the learner.*
- *Provide elaborated feedback to enhance learning.*
- *Present elaborated feedback in manageable units.*
- *Be specific and clear with feedback message.*
- *Keep feedback as simple as possible but no simpler (based on learner needs and instructional constraints).*
- *Reduce uncertainty between performance and goals.*
- *Give unbiased, objective feedback, written or via computer.*
- *Promote a "learning" goal orientation via feedback.*
- *Provide feedback after learners have attempted a solution.*

Things to avoid:

- *Do not give normative comparisons.*
- *Be cautious about providing overall grades.*
- *Do not present feedback that discourages the learner or threatens the learner's self-esteem.*
- *Use "praise" sparingly, if at all.*
- *Try to avoid delivering feedback orally.*
- *Do not interrupt learner with feedback if the learner is actively engaged.*
- *Avoid using progressive hints that always terminate with the correct answer.*
- *Do not limit the mode of feedback presentation to text.*
- *Minimize use of extensive error analyses and diagnosis.*

The framework presented here differentiates between three types of feedback:

1. *Ingame feedback*: This type of feedback refers to implicit feedback during the role-play session through the reactions of the chatbot, which can be verbal or non-verbal (e.g., facial expressions). This feedback is important for simulating authentic real-life situations and helps the learners to understand the consequences of their actions.
2. *Aftergame feedback (general)*: This type of feedback describes a general summary of analysis results, which provides overall feedback on the learners' performance during the role play session, summarizing the most important aspects (positive and negative).
3. *Aftergame feedback (specific feedback on single actions)*: This type of feedback refers to direct and specific feedback on single incidents during the role play session provided through prompts in an interactive replay of the conversation. In this replay, the whole chat conversation is shown again step by step, augmented and enriched with individual feedback on specific actions of the learner. It is possible to navigate between different phases of the conversation in the replay, pause it, jump to the next feedback marker, or even search for certain keywords. These properties make the interactive replay much more flexible and searchable than, for example, conventional role play videos.

Additional types of feedback are possible. Conceivable in collaborative environments could be, for example, a direct exchange between players (e.g., in form of a chat) as an opportunity to

give each other feedback. Also imaginable are group reflection sessions administered by an expert with decided feedback on the participants' performance (see paragraph *Immersion and reflection*).

In summary, three conceptual features characterize the framework (from the design perspective): (1) chat-like interaction with AI-controlled chatbots, (2) the separation of phases of immersion (role-playing) and reflection, and (3) adaptive feedback based on individual performance analyses. Based on these components, several different scenarios and use cases can be developed.

1.2.1.2 *Technical Approach*

On the technical level of the proposed framework, the implementation of serious role-playing games for the training of social skills entails three main challenges: dialog modeling of the chatbots, implementation of a multi-agent blackboard system as the backbone to keep components independent from each other and support performance optimization due to parallel processing, and automated performance analysis and feedback generation. These three components will be illustrated in the following paragraphs:

Dialog modeling. As mentioned previously, this framework uses AIML for implementing the chatbots' conversational logic. AIML follows a simple pattern matching algorithm but offers a variety of elements that make it much more than a simple database of questions and answers. These elements include (Wallace, 2004):

- *Recursion*: In AIML, the `<srai>` operator is used to implement recursion. There are various applications for this operator, e.g., *symbolic reduction* to reduce complex grammatical forms to simpler ones, splitting an input into two or more subparts and combine responses to each of them (*divide and conquer*), detecting keywords anywhere in the input, covering synonyms and spelling or grammar mistakes.
- *Context*: The AIML element `<that>` refers to the chatbot's previous utterance, which enables it to respond based on the context. Remembering the last utterance is especially important if the chatbot asks a question. If the user provides an answer to a question, the chatbot needs to know to which questions the answers are referring to, otherwise it cannot provide a meaningful answer. Another option to preserve context in AIML is the `<topic>` tag. It appears outside a category and bundles them into general topics. This way, the chatbot knows the current subject area and can generate specific responses corresponding to the current topic.
- *Conditionals*: In AIML, it is possible to write conditional branches using the `<condition>` tag. This is done by checking the value of a variable and returning a response depending on that value. Variables can be either predefined (e.g., name, gender, location) or self-created. `<get>` and `<set>` tags can be used to manipulate and retrieve the current content of the variables.

Although AIML has a long history and constitutes a common solution for AI-controlled chatbots in educational contexts, it has several limitations. Some general workarounds for overcoming these limitations as well as tricks for the creation of believable and authentic

chatbots have already been illustrated in Chapter 1.2.1.1, but on the implementation level, the framework proposes some concrete technical workarounds and solutions:

- *Sentence openers*: A central element is the already mentioned use of sentence openers. This strategy simplifies the input analysis in many ways and thus offers several advantages (see paragraph *Chat-like interaction with AI-controlled chatbots*). AIML chatbots cannot truly grasp the sense of what has been said, but by using sentence openers, the general gist of a message is predefined (e.g., affirmation, disaffirmation, or further inquiry), which simplifies the interpretation enormously. This way, it is at least possible to provide a default answer tailored to the specific sentence opener even if no predefined pattern matches the free text input following the opener. Even more context for the chatbot can be provided by dividing the conversation into several phases and having unique sentence openers in each of them. This way, the use of sentence openers helps to reduce the complexity of the dialog scripts drastically, as all possible input sentence starters are known. Of course, creating elaborate and sophisticated AIML scripts still requires high developmental effort, but the scope of this task can be greatly reduced. In addition, using sentence openers makes it possible to distinguish between different users in a collaborative environment, because otherwise it is impossible for the chatbot to know who is speaking. By introducing slightly different sentence openers for different users, this problem can be worked around.
- *External triggers*: AIML chatbots are passive, they only react to an input they receive. They cannot take the initiative. By using external triggers, this behavior can be bypassed. The triggers can make the bot become active when this is required in certain situations.
- *Feedback tags*: The introduction of feedback tags, which are added to the learners' input by analysis agents, enables the creation of feedback that is related to specific points of the conversation and can be provided to the learners' during the replay within the post-role-play reflection phase. Through the tags, these situations can be marked. They may, for example, include `#praise#` for especially positive contributions, `#interruption#` for situations in which learners interrupt other persons, `#repetition#` when learners send the same message repetitively, and many others depending on the concrete context and task within the game (see paragraph *Performance Analysis and Feedback Generation*).

Multi-agent architecture. The technical framework is essentially built on a multi-agent system architecture with a blackboard as the communication and integration mechanism, which is realized through a *tuple space*. In this system, components (called "agents") are loosely coupled, which means they do not communicate with each other directly, but through exchanging entries (called "tuples") on a central tuple space server (Gelernter, 1985). Tuples have a simple structure containing primitive data types (integers, characters, booleans) and strings. In a tuple space, several communicating processes can use tuples to exchange data, i.e., sending tuples to the tuple space or withdrawing tuples from it. Each tuple has an independent existence in the tuple space and is equally accessible to all processes in it (Gelernter, 1985). According to the original concept of Gelernter, there are only a few generic

operations (read, write, take, wait-to-take, etc.) to interact with the tuple space, but there are also active trigger mechanisms such as notifications. The specific implementation used in the framework presented here is called SQLSpaces, which has been developed in the COLLIDE group[3] of the University of Duisburg-Essen (Weinbrenner, 2012). SQLSpaces works on a relational database and translates operations into SQL statements. There are several advantages to SQLSpaces over other tuple space implementations: (1) persistency provided by the underlying database, (2) a versioning system, (3) awareness features, and (4) extendibility towards other systems (Malzahn, Weinbrenner, Hüsken, Ziegler, & Hoppe, 2007). The server itself is implemented in Java, but SQLSpaces system framework provides clients of the agent programming in various programming languages. Despite serving as the main communication hub, SQLSpaces also facilitates the logging of relevant data of each gaming session, which can later be used for analysis and comparison.

The overall system architecture includes a client (user interface) and several agents, each responsible for exactly one task in either game control, dialog analysis, or feedback creation. All training scenarios implemented based on the framework have been realized as web applications using HTML, CSS, and JavaScript (2D frontend). Previous implementations were based on OpenSimulator3 (3D frontend). The decision to turn away from the 3D approach was based on the fact that no specific advantages of 3D environments over 2D environments could be shown at that time (Malzahn, Buhmes, Ziebarth, & Hoppe, 2010). The general system architecture is shown in Figure 1.

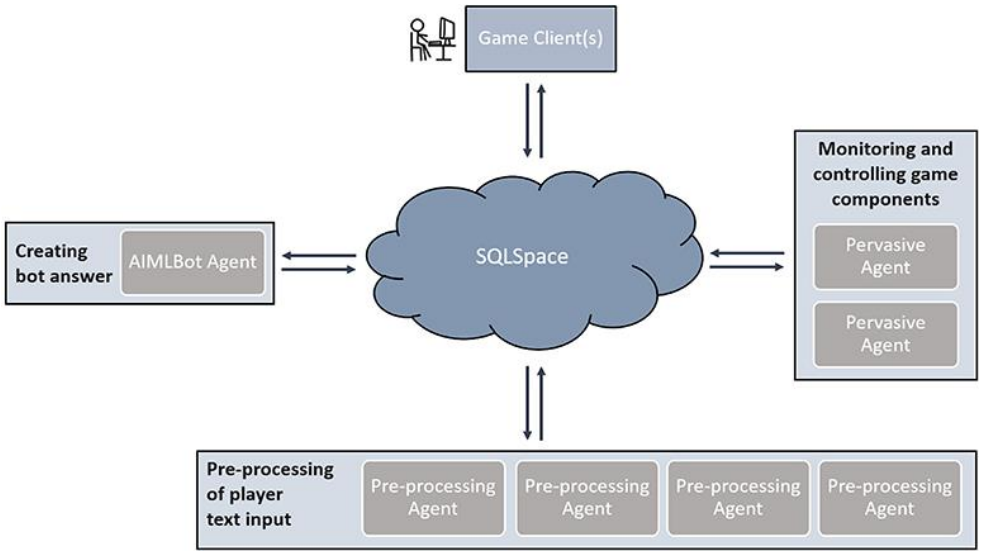

Figure 1: Basic system architecture (Othlinghaus-Wulhorst & Hoppe, A Technical and Conceptual Framework for Serious Role-Playing Games in the Area of Social Skill Training, 2020)

The client (user interface) and the agents are writing, reading, and withdrawing tuples from the tuple space without communicating with each other directly, thus making it a loosely coupled and adaptive system, which means that agents can – depending on the actual application scenario – easily be added, removed, or modified. All agents can read/ write tuples from/into the tuple space. They are waiting for a tuple to appear that has exactly the structure they can process and use the information provided by the read tuple to enrich the data with

[3] https://collide.info/

their analysis results and then write a new tuple to which other agents might react. Because every agent can only process a special form of tuple, a certain direction is set, which can be considered the *game loop*.

There are three different groups of agents, depending on their functionality and task within the system: pervasive agents, pre-processing agents, and the AIMLBot agent. *Pervasive agents* are responsible for the game control. They are overarching agents that connect the individual game components with each other and assume tasks like managing the log-ins, arranging a pair of players (in a collaborative scenario), and checking if players have been inactive for a certain amount of time. *Pre-processing agents* are pre-processing the player's input before the chatbot's answer is generated. The pre-processing is needed to generate the best possible answer by overcoming the limited capabilities of AIML. The separate analysis of important conversational aspects helps to prioritize specific behaviors. This procedure ensures that the chatbot reacts adequately to, e.g., rude or aggressive behavior of the player(s). Apart from that, it reduces the complexity and size of the AIML scripts, supports the feedback creation, and allows to build up a score system. Each of the pre-processing agents analyzes the user's input with regard to one specific aspect (e.g., rude or aggressive behavior, use of emoticons, or especially polite or helpful behavior). Depending on the context and concrete scenario, different pre-processing agents need to be used. The concrete functionality of the agents depends on their actual task, e.g., they compare the player's input with pre-defined wordlists to search for specific keywords or phrases or calculate specific values like writing speed or inactivity. The following list includes examples of agents and their respective tasks (Emmerich, Neuwald, Othlinghaus, Ziebarth, & Hoppe, 2012) (Othlinghaus-Wulhorst, Mainz, & Hoppe, 2019):

Pervasive agents:

- *Register agent*: The register agent is managing the log-ins. When a new client logs in, the agent receives a request (callback) via the tuple space and initiates a new game session. In the process, information about the player (such as name, role, and level) to check if this player has already played the game before and, based on the existing data, select the appropriate level and/or role (depending on the concrete scenario and game modality).
- *Silence agent*: The silence agent is based on timers (one per session). These timers are reset when a new conversation tuple has been written into the tuple space. If the player is not sending any messages after a certain amount of time, it evokes an event triggering a chat bot message reminding the player to answer (and even cancels the conversation after several reminders).
- *Collecting agent*: The collecting agent collects the results of the pre-processing agents and summarize their findings in a new tuple, which signalizes that the preprocessing has been completed and hence activates a call back in the AIMLBot agent.
- *Score agent*: The score agent is calculating a general score reflection of the player's performance during the conversation. Aggressive or rude behavior as well as the use of forbidden words or phrases have a negative effect on the score. Contrariwise, especially polite behavior is positively noted. In addition, different answer qualities (unhelpful, neutral, and helpful) have been defined to determine whether a chosen

answer moves the conversation forward (positive rating) or fails to do so (negative rating). Furthermore, the total number of messages, message time (the time a player needs for sending a message after receiving a chat bot answer), and writing speed (the time it took to send a message divided by the message length including the sentence opener) influence the score.

- *Serialization agent*: The serialization agent is responsible for logging and saving all game data (e.g. in an XML file). The saved protocols can later be used for external analyses.

Pre-processing agents:

- *Aggression agent*: The aggression agent checks whether a player is being aggressive. First, this agent compares the player's input to a pre-defined list of expressions related to threats and violence. Second, the agent searches for consecutive exclamation marks and words or phrases written solely in capital letters (capslock). Both variants are as interpreted as screaming according to common chat custom.
- *Rudeness agent*: The rudeness agent searches for swearwords and defamations in the player's input by comparing it to a list of related pre-defined keywords. It saves information on whether and how often a player is using abusive language during the conversation.
- *Politeness agent*: The politeness agent searches for especially polite keywords and phrases by comparing the player's input against a corresponding pre-defined list.
- *Step agent*: The step agent is counting the overall number of messages, which is relevant for the score.
- *NoGo agent*: The NoGo agent is searching for phrases that should never be used even if they are neither aggressive nor rude. The actual content depends on the concrete scenario and task.

All pre-processing results are collected, and if an immediate reaction to a specific behavior of the learner is required, the text input is modified accordingly. If this is not the case, the AIMLBot agent receives the original text input. The AIMLBot generates an answer based on the underlying AIML script(s) and finally writes a new tuple into the tuple space, containing all information that was collected so far plus the bot answer. This answer may also include any of the above-mentioned feedback tags in order to mark this message as an important point for later feedback. The final tuple provided by the AIMLBot agent is read by the game client(s), which display(s) the player's message and the bot's answer on the screen (Emmerich, Neuwald, Othlinghaus, Ziebarth, & Hoppe, 2012).

Performance analysis and feedback generation. Performance analysis and feedback generation are important factors for successful learning. Of course, both always depend on the concrete learning objectives and context of the specific game scenario. Feedback allows the learners to reflect on their actions during the role-playing session, to analyze the consequences, and to assess their behavior.

As described in the previous passage, the system architecture used in this framework is applying analysis agents for evaluating the player performance. Each of them is responsible for the evaluation of one specific aspect of the player's communicative behavior and actions.

All analysis agents evaluate the user's message and add their feedback in form of feedback tags such as:

- *#praise#*: the player's contribution to the conversation is positive
- *#interruption#*: the player is speaking without being asked to, interrupting his dialog partner(s)
- *#repetition#*: the player is sending the same message twice (or even more often)
- *#criticize#*: the player's contribution does not comply with the previous question/statement of a dialog partner (e.g., if an open question is answered with yes or no)
- *#cancel#*: the dialog partner is cancelling the conversation as no compromise solution can be found
- *#rejection#*: the player refuses to talk about a certain topic or to agree to a proposed solution
- *#rudeness#*: the player's contribution contains rude content
- *#aggression#*: the player's contribution shows aggressive behavior
- *#politeness#*: the player's contribution is especially polite
- *#silence#*: the player has been inactive for a certain amount of time

These tags are just examples; it depends on the concrete scenario which ones are actually used in the respective game. The feedback tags that are generated during the role-playing session are not displayed in the actual conversation during the role-playing session (so players do not get to see them while they are playing the game), but can later be used to provide a comprehensive and detailed feedback to the player, both on an overall level and in form of an interactive replay enriched with situational explanatory annotations (see paragraph *Adaptive Feedback*). Figure 2 shows an example of a replay of the role-playing session in ColCoMa, one of the implemented case studies:

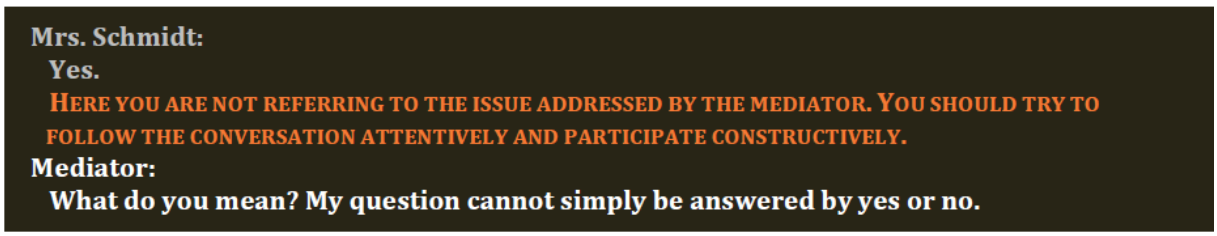

Figure 2: Augmented replay during the reflection phase in the game ColCoMa (Emmerich, Neuwald, Othlinghaus, Ziebarth, & Hoppe, 2012)

Appropriate ingame feedback (implicit feedback during the role-play session through the reactions of the chatbot) is ensured through a pattern matching hierarchy provided by the pre-processing agents. This allows situational and appropriate reactions of the chatbot. If one of the agents detects something that requires an immediate reaction of the chatbot (e.g., rude or aggressive behavior), the input string of the player is replaced by a specific trigger (e.g.,

"swearword"), which is causing the bot to react appropriately. The chatbot is only receiving the original text input of the player in case no pre-processing agent is raising an alarm.

In summary, the technical conception is based on three main components: (1) AI-controlled chatbots that adapt to the player's behavior, (2) a multi-agent blackboard system as the backbone, and (3) intelligent support for automated performance evaluation and feedback generation. This setup results in a loosely coupled and efficient system composed of independent components and facilitates the creation of serious role-playing games in that it allows for tailoring and adapting the given core architecture to any new scenario with very limited effort.

1.2.2 Case Studies

Two major case studies have been developed within the scope of this thesis based on the framework introduced in the previous chapter (and elaborated on in detail in Chapter 2). The two training scenarios used in these case studies include workplace-oriented conflict management (game: "ColCoMa"; Chapter 3) and customer complaint management (game: "CuCoMag"; Chapters 4–5).

1.2.2.1 *Case Study 1: Conflict Management*

The first case study, which is presented in Chapter 3, is based on a scenario designed for the training of conflict management skills in a workplace-oriented context. The game *ColCoMa* (Collaborative Conflict Management) is a collaborative serious role-playing game involving two human players and a chatbot in a 2D virtual environment. Apart from the main publication presented in Chapter 2 (Othlinghaus-Wulhorst, Jedich, Hoppe, & Harrer, 2018), the following summary is referencing Emmerich, Neuwald, Othlinghaus, Ziebarth & Hoppe (2012).

Context. Conflict management is considered a pedagogically relevant application field of role play in general and serious role-playing games in particular. Conflicts are part of everyday work life and pose a topic that concerns many companies. Handling a conflict in a constructive way – especially in a work environment – has become an increasingly important social skill (Erdmüller & Jiranek, 2010) and is subject of professional training (Emmerich, Neuwald, Othlinghaus, Ziebarth, & Hoppe, 2012). Virtual role-play scenarios provide a good opportunity to try out, experience, and learn the aspects of conflict management without putting oneself in conflict situations and having to bear serious consequences (Cheong, et al., 2011). Additionally, collaborative learning has great potential as it can improve social competencies and skills (Turani & Calvo, 2006) (Emmerich, Neuwald, Othlinghaus, Ziebarth, & Hoppe, 2012). Furthermore, the use of chatbots in this context is advantageous because a high degree of standardization can be achieved.

Approach. In this game, a pair of players has a conversation about a given fictitious conflict within a graphically represented virtual chat setting. The conversation is moderated by a chatbot acting as a mediator, while the human players are taking on the roles of the two conflicting parties. The main goal in this game is to come to a conflict resolution by showing appropriate and constructive behavior during the conversation. It follows the typical structure of mediation talks. Each of the two players is assigned a pre-defined role in the conflict

scenario: Mr. Meier is a member of the computer support hotline team in a fictitious software company and often takes much time for his customers. Mrs. Schmidt is his supervisor and does not endorse the long call sessions Mr. Meier has with the customers, because he keeps other customers on hold. She wants him to work more efficiently and expresses this through a negative appraisal of Mr. Meier's performance. As a result, the situation escalates. The scenario is preferably kept simple and comprehensible in order to support immediate understanding and empathy with the assigned role.

Gameplay. The background story illustrating the conflict and its causes are conveyed by an introduction in form of a cartoon-like picture story, whereas both players receive different illustrations – each from the perspective of the respective player's role, which is supposed to result in conflicting points of view. During the actual role-playing session, i.e., the mediation talk, the players see a cartoon-like graphical representation of the two dialog partners (the other player and the mediator) on screen as if they would sit opposite the dialog partners. The players can communicate with each other and the mediator by means of an integrated chat using pre-defined sentence openers, which need to be supplemented with free text input (see Chapters 1.2.1.1 and 1.2.1.2 for detailed information about the functionality of sentence openers in chat-based serious role-playing games). The players can also evoke facial animations using common emoticons. The set of sentence openers offered to the players changes several times during the game. It always depends on the current state of the conversation. In this specific scenario, the sentence opener also serve another purpose: In order to allow the mediator chatbot to differentiate between the two players, each player is provided with slightly different sentence openers. Otherwise, the chatbot would not know which player is talking at any given point in time, which is crucial for the course of the conversation as well as for the individual feedback generated for each player. In addition to the chat area, the players have a notepad and help section at their disposal to get additional information about the scenario, their role, game controls, and possible actions. The conversation follows the typical structure of mediation talks:

1. *Framing phase*: In this introductory phase, rules for the participants' behavior towards each other throughout the conversation need to be established, whereas the actual context is not yet the focus. The conflicting parties are asked to describe their personal hopes and goals for the mediation, which is meant to help them to reflect on their own positions and understand the opponent's point of view.
2. *Topic collection*: In this phase, the conflicting parties are supposed to name topics they would like to discuss during the mediation talk, without commenting on them individually. In the given scenario, possible topics are, for example, the performance review, working conditions, the interaction with each other, and future perspectives.
3. *Working on the conflict*: This phase is the actual core of the mediation talk. The participants get to discuss the collected topics in detail. Both participants get the chance to say why a certain topic is important to them, what they would like to change about it, and what each of them could do to achieve this. The participants are not only asked to describe their own point of view, but also how they perceive the other party's perspective.

4. *Looking for a solution*: The task during this phase is to find solutions for each of the topics all parties can agree with.
5. *Contract*: If satisfying solutions could be found, the conflicting parties make a virtual contract.

There are two possible outcomes to this scenario: Either the conversation is successfully finished by achieving a conflict resolution or it is cancelled in case the mediator chatbot notices that the conversation is stuck or one of the players leaves the conversation.

Post-role-play reflection. Irrespective of how the mediation talk ends, it is followed by a separated reflection phase in which participants are supposed to step out of their role and reflect on their behavior from a third-person perspective (see Chapter 1.2.1.1 paragraph *Immersion and reflection*). In a first step, the two players have the opportunity to directly exchange feedback with each other by means of a free chat without the mediator. In a second step, they receive a recapitulatory textual feedback on their overall performance during the conflict conversation. This is followed by a replay session in which the whole conversation is recaptured step by step. The replay is augmented with detailed individual feedback at certain points of the conversation with respect to the respective player's performance, commenting on especially positive (e.g., constructive suggestions, adhering to rules, picking up suggestions from the opposite party) and negative (e.g., interrupting the dialog partners, offensive behavior, refusal) contributions. Performance evaluation is based on general rules conflicting parties must abide by during mediation talks. These rules include, for example, no swearing, not being rude, aggressive, or reproachful, and not impairing the autonomy of the dialog partner (Stauss & Seidel, 2010). Instead, they are supposed to show an open and constructive attitude, help their counterpart to understand their perspective, and name issues and topics in a concrete way (Stauss & Seidel, 2010). The analysis agents described in Chapter 1.2.1.2 have been adjusted to match these criteria and provide an exact assessment of the specified behaviors. A detailed description of the functionality of the different analysis agents and the concrete operationalization of the described parameters can be found in Chapters 2 and 3.

Research questions and hypotheses. Apart from assessing the general game experience and perception of the prototype from the players' point of view, the preliminary focus of the evaluation study presented in Chapter 3 was on the collaborative aspect of the game. The main goal of the mixed-method study was to investigate the question if there is a correspondence between gaze synchronicity of the players and the quality of collaboration, which was analyzed with help of eye-tracking. The standard assumption is that a certain extent of convergence of the visual foci of attention between players in a collaborative or cooperative scenario indicates better coordination and consideration of the other party, whereas this assumption has been refined by taking into account the different roles (including the one of the chatbot). Concretely, the supposition was that there would be:

1. A positive relation between the convergence of visual foci (i.e., gaze synchronicity) of the two human players and the successful completion of the game (hypothesis 1)
2. A positive relation between the convergence of visual foci and the quality of collaboration in the chat (hypothesis 2)

3. A dynamic (time-related) congruence between similar eye movements (synchronicity) and the quality of collaboration in the chat (hypothesis 3)

In this study, *gaze synchronicity* has been defined as the extent to which the two players of a gaming session have been looking at the same areas of interest within the same time interval during the session. The parameters *success in the game* and *quality of collaboration* have been operationalized based on several complex criteria (see paragraph *Method of analysis*).

Experimental Design. Twenty subjects (average 22.8, SD = 2.84, 5 females, 15 males) participated in the study and have been tested in dyads. The players' gaze has been tracked during the experiment with the help of two desktop-based eye-trackers. The role distribution was randomized. There were two possible outcomes in the given scenario: Either the participants completed the conversation successfully by reaching a conflict resolution or it has been cancelled by either the mediator chatbot or one of the players. The gaming session was followed by answering several post-experiment questionnaires.

Method of analysis. In order to be able to compare the gaze movements of the two players of each dyad, the chat interface of the game has been subdivided into six areas of interest (three sections of the chat protocol, input area, mediator face, and partner face). The eye-tracking data of the two players of a dyad have been compared by matching the same area of interest at a specific time stamp (the detailed calculation is explained in Chapter 3). The matches were finally summed up and divided through the total number of seconds to normalize against varying experiment duration, resulting in a percentage that indicates to what extent the two players have been looking at the same areas of interest at the same point in time during the course of the game. This percentage is called *convergence of visual foci*. The success in the game has been measured by developing a so-called *achievement score*, reflecting the players' performance during the mediation talk based on three criteria:

1. Automated feedback generated by the system (summarizing the players' behavior during the conversation)
2. The successful completion of the topic collection phase (which is considered to be a milestone within the game)
3. The achievement of a final conflict resolution at the end of the role-play (by verbally signing a contract between the two conflicting parties, which puts the arrangements and rules the dyad worked out together with the mediator into writing)

Collaboration quality was rated based on a five-dimensional rating scheme:

1. Argumentation (players bring forward or discuss arguments)
2. Agreement/disagreement (players endorse or dissent from one another)
3. Collaborative orientation (players refer to each other, ask questions, or provide feedback)
4. Solution orientation (players try to find a solution)
5. Shared awareness/reinforcing shared history (players share common knowledge or explain their respective situation)

The whole transcript of each gaming session has been analyzed and checked against the five dimensions and rated with a total quality score. Similar to the achievement score, all matches

of the session are added up to a percentage, which indicates the overall collaboration quality for a dyad.

Evaluation results. The results of the study confirm a correspondence between the convergence of visual foci of attention and the success in the game. It appears to be plausible that the outcome of the game is influenced by the extent of gaze synchronicity and not vice versa, i.e. the more similar the visual foci of attention are, the better players collaborate with each other (although a casual relation cannot be verified). The results of the study also confirm a positive relation between the convergence of visual foci of attention and the quality of collaboration, especially in regard to the three dimensions agreement/disagreement, solution orientation and shared awareness. Thus, it can be assumed that a higher gaze synchronicity results in a higher collaboration quality. Surprisingly, only strong correlations on the aggregate level could be found (taking overall eye-tracking convergence as a global parameter), but not in terms of synchronicity between convergent eye-tracking and chat interaction. In can be argued that the specific nature of the game environment is responsible for these results, since in total three persons are involved in the conversation and the two human players to not communicate with each other directly. Due to the specific nature of AIML and the dialog modeling, the players always talk to the mediator chatbot and reply to his questions. Thus, they can never really react to the other conflict party's messages immediately. This means that the structure of the conversation, is pre-defined to a certain extent and also the time interval between the messages of the dialog partners is rather high, because the mediator is always writing in-between. All in all, the results of the study indicate that a certain convergence of the visual foci of attention between cooperating partners indicate higher success and collaboration quality, although only on an aggregate level. No generalization can be made due to the small amount of participants, but the results are promising.

1.2.2.2 *Case Study 2: Customer Complaint Management*

The second case study is based on a scenario designed for the training of customer complaint management. The game *CuCoMaG* (Customer Complaint Management Group Reflection) is a serious role-playing game based on theories of consumer psychology and complaint management, originally developed in 2016 (Othlinghaus & Hoppe, 2016) (Doberstein, et al., 2016) and re-designed and evaluated in 2019 (Othlinghaus-Wulhorst, Mainz, & Hoppe, 2019). It involves a human player (assuming the role of a customer service employee) and a chatbot (in the role of a complaining customer is reporting a certain problem).

Context. Online shopping has become indispensable in Germany and many other countries all over the world. Studies show that it is more popular than ever before. 66,4 percent of Germans regularly shopped online in 2019, with the fashion sector accounting for a particularly relevant share of around 30 percent (Handelsverband Deutschland – HDE e.V., 2020). The rising demand in online retail can lead to increased customer complaints, which in times of digitalization are increasingly voiced web-based. While proper customer complaint management offers retailers the opportunity to strengthen customer loyalty and prevent customer churn, it can be challenging for untrained employees. They are challenged to find a fair solution, put themselves in the customer's shoes, while always remaining friendly and calm (Stauss & Seidel, 2014). In order to prepare employees for different types of customers

and to expand their competence in dealing with customers, training in complaint management is useful. Video-based learning and role plays are effective tools for conveying the proper concepts and develop professional behavior towards complaining customers (Heung & Lam, 2003). Virtual role plays in particular are a great medium to help people experience conflict situations and try different problem-solving strategies. They can learn how to act and re-act adequately in these situations without having to fear consequences in the real world.

Approach: In the scenarios included in CuCoMaG, the player takes on the role of an employee of the fictitious company "LittleONEs". The company sells personalized children's clothing via an online shop. An online shop in the fashion sector is particularly suitable for complaint management, since both accessories and clothing are the product categories most associated with complaints (Cho, Im, Hiltz, & Fjermestad, 2002). In addition, there are many competitors with high quality elasticity in the market to which dissatisfied customers might migrate, making successful complaint management particularly important (Fornell & Wernerfelt, 1988). Moreover, since children's clothing does not require complex warranty regulations, customers have much leeway for individual complaints.

Gameplay. In CuCoMaG, the player communicates with the customer chatbot through a simple chat environment in the style of standard online customer support tools available on the market. Just as in ColCoMa, the player has to select a sentence opener from a pre-defined set and amend it with free text input in order to create a chat message. Here, too, the displayed sentence openers are dependent on the current state of the conversation. In addition, the player can also access the company's database to search for additional information about the respective customer and the order details. This is necessary for the player to get all information required to resolve the situation. CuCoMaG entails three different scenarios, which mainly differ based on the type of customer used (in terms of conflict style) and the problem situation of the respective customer, which results in different levels of difficulty. In general, complainants can be distinguished by their conflict management style, which is based on their individual character and approach to customer service. With their negotiation model "five styles of handling interpersonal conflict", Rahim and Bonoma defined five basic styles of conflict management that influence the behavior towards the customer service employee (Rahim & Bonoma, 1979):

1. *Integrating customer*: the customer is willing to exchange information in order to find a solution that is acceptable to both parties.
2. *Obliging customer*: the customer tries to downplay the differences and emphasize the similarities in order to satisfy the other party's request.
3. *Dominating customer*: the customer does everything to achieve his goal and therefore often ignores the needs and expectations of the other party.
4. *Avoiding customer*: the customer avoids confrontation and does not manage to satisfy either his own concerns or those of the other party.
5. *Compromising customer*: both parties give up something in order to reach a mutually acceptable decision.

In addition to these negotiation styles, there are three special types of complainants that require special attention (Stauss & Seidel, 2014): If a customer complains about the same

problem several times within a certain period of time, this is called the repeat or *multiple complainant*. *Follow-up complainants* are customers who contact the company again during or after the processing of their complaint because they are dissatisfied with the complaint processing or the complaint result. If the customer has no comprehensible reason for his complaint, makes disproportionate demands, or even wants to harm the company, he can be classified as a *grouser*. All these types require special treatment in customer service and are represented in CuCoMaG. The specific problem situations of the customers included in the game are based on a study conducted by Cho et al., who investigated current causes and sources of online complaints (Cho, Im, Hiltz, & Fjermestad, 2002). Table 1 summarizes the attributes of the three different scenarios in CuCoMaG:

	Scenario 1	**Scenario 2**	**Scenario 3**
Level of difficulty	Easy	Medium	Hard
Type of customer	Integrating	Compromising (follow-up complainant)	Dominating (grouser)
Reason for complaint	Delivery problems	Problems with the customer service	Problems with the business terms and conditions
Goal	- Tutorial/introduction to the game and the user interface - Experience the basic milestones of a complaint conversation	- Apply soft skills - Pass through all five phases of a complaint process successfully	- Learn to deal with extreme situations and provocations - Apply *active farewell* (Stauss & Seidel, 2014)

Table 1: Scenarios in CuCoMaG

Each scenario is sub-divided into five different conversational phases representing the typical structure of complaint conversations according to Stauss and Seidel (Stauss & Seidel, 2014):

1. *Greeting phase*: In this first phase, the two dialog partners introduce themselves and the complaining customer is given the opportunity to briefly describe the problem.
2. *Aggression-reduction phase*: In this phase, the customer should have the opportunity to explain the situation from his/her perspective and to vent his/her anger. It is possible that the customer attacks the support employee in this phase.
3. *Conflict-settlement phase*: During this phase, the support employee's task is to direct the conversation to an objective level and to ask concrete questions in order to gain detailed information that are important for processing the issue.
4. *Problem-solution phase*: In this phase, possible over-expectations of the customer are to be reduced. A meaningful problem solution is to be found, which should be justified by the support employee in any case, in order to establish understanding of the customer.
5. *Conclusive phase*: This is the conclusive phase, in which the support employee must ensure that the customer has understood and accepted the final solution. The fare well should end with a positive formulation.

Two main dimensions determine the course of the game and are demanded of the player: On the one hand, the player has to complete the information milestones. It is his task to obtain enough information to select a suitable compensation service. This is particularly decisive in the greeting phase, the conflict-settlement phase, and the conclusive phase (Stauss & Seidel, 2014) (Goodwin & Ross, 1990). One the other hand, the player has to prove soft skills to reassure the customer and generate sympathy. Soft skills are particularly important in the greeting phase, the aggression-reduction phase, and the problem-solution phase (Stauss & Seidel, 2014) (Goodwin & Ross, 1990), but they actually pervade the entire complaint conversation. The communication during the whole process of complaint handling is very important to the success of restoring customer satisfaction (Stauss & Seidel, 2014). Particularly important soft skills related to customer complaint management are communication skills, conflict management, and empathy. Content-related aspects that must be taken into account are initial wording, problem repetition, conflict settlement, problem solution, and concluding wording (Stauss & Seidel, 2014). There are three possible outcomes for each of the scenarios: (1) The player completes the complaint conversation successfully, (2) the player leaves the conversation, or (3) the player is inactive for a certain amount of time and fails to react to repeated requests of the customer, which will cause the customer chatbot to terminate the conversation.

Post-role-play reflection. Irrespective of how the complaint conversation ends, the actual role-play phase is (possibly time-delayed) followed by a reflection phase. In contrast to ColCoMa, CuCoMaG offers explicit support for group reflection, meaning several participants who individually played the game come together to reflect on and discuss their performance together with an expert who moderates the session. There are numerous reasons why group reflection can be advantageous. Generally, people have a consciousness of themselves (Greif, 2008). They are able to observe their own actions and reflect on themselves. According to Eckensberger, people are potentially self-reflective subjects (Eckensberger, 1998). However, not only are people capable of self-reflection, but they are also capable of reflection in groups to which they feel they belong. The need is strong to share and compare reflections with others. In addition, reflections open access to higher forms of learning and conscious change in the individual's or group's actions (Greif, 2008). Individual self-reflection is a conscious process in which individuals reflect on and address their ideas and actions related to the real and ideal self-concept. In addition, outcome-based self-reflection describes the process in which the person not only consciously reflects on herself, but also develops inferences for future actions or self-reflections in the process. Deliberate self-reflection requires the activation of intuitive self-awareness. Outcome-oriented group self-reflection is the process by which members of a group talk about the group self-concept or about actions of individual group members and common actions that relate to the real or ideal group self-concept. This process of information processing requires linguistic communication but can also function through non-linguistic means. Prerequisites for group reflection are the activation of prior or contemporaneous individual self-awareness and reflection processes within the group (Greif, 2008). There are numerous factors that can support but also disrupt self and group reflections. Communication can be a possible trigger and supporting factor for reflection (Greif, 2008). In a study by Spafford and Haarhoff (2015), intervision was found to provide better containment for most students and broaden their perspectives (Spafford & Haarhoff, 2015). Based on these

findings, intervision is suggested as a useful approach to using group reflection. In addition to the potential factors, there are also numerous advantages and disadvantages. For example, self-reflection can increase motivation to reduce discrepancies between the real and ideal self-concept, which can also have a negative effect if it causes the person to confront his or her own negative characteristics (Greif, 2008). Another study by Kim et al. (2009) demonstrates a greater increase in knowledge through group reflection than through individual reflection (Kim, Hong, Bonk, & Lim, 2009). CuCoMaG features a separate group reflection tool designed for supporting the group reflection phase. It has been developed for use in a collaborative training center environment. It visualizes the performance of the participants during the game in a dashboard design. An expert or trainer is supposed to lead the discussion, and the tool allows him or her to present and compare the performances and results of different players. Due to the fact that the actual gaming session is separated from the reflection phase, the trainer can take some time before meeting with the participants to review the different chat transcripts, make notes about anything he or she wants to talk about, and compare certain characteristics or sequences. The group reflection tool includes several components to analyze and visualize important performance data:

- *List of participants*: The players who participate in the group reflection session are listed with their achieved scores. The trainer can load an unlimited number of player files into the system.
- *Charts*: Different charts can be selected for visualizing the players' performances during the conversation with the customer chatbot. A bar chart shows how many times certain selectable characteristics have been used by each player (e.g., especially polite, rude, or aggressive behavior). A line chart displays the players' performance (score) over time. Trainers can directly switch from the line chart to a certain point of interest in the chat history by simply clicking on a point in the chart.
- *Replay*: For each participant of the current group reflection session, a replay of his or her chat conversation(s) is available – enriched with annotations generated by the system based on the individual performance analysis.
- *Notepad*: A notepad offers the possibility to make notes about each participant. Chat messages from the replays can be directly copied into the notepad and annotated.
- *Report*: A report for each participant can be created from the notepad, which can be useful for further consideration and progress.

The group reflection tool has not been part of the study presented in Chapter 5, but it has been re-designed and evaluated in 2020. The additional (so far unpublished) results of this follow-up study are described in Chapter 6.1.

Research questions and hypotheses. The main goal of the mixed-method evaluation study presented in Chapter 5 was to investigate whether the developed scenarios and chatbots used in them qualify for a real training situation. Here again, qualitative and quantitative methods have been used to gain insights regarding the playability of the prototype and the perception of the chatbots from the users' point of view. It was assumed that the game experience and perception of the chatbots for participants who played the second scenario (compromising

customer) differed from the experience of those who played the third scenario (dominating customer). Specifically, it was assumed that the results differ in the dimensions *tension*, *negative affect*, and *challenge* of the GEQ[4] (hypothesis 1). Furthermore, the assumption was that the results differ in the dimensions *comfortable*, *thoughtful*, *polite*, *responsive*, and *engaging* (hypothesis 2). In addition, it was assumed that participants with prior experience in complaint management perform better than participants without prior experience.

Experimental Design. 20 participants (average 26.05, SD = 7.99, 15 females, 5 males) participated in the evaluation study; 4 of those participants indicated that they had prior experience in customer complaint management. After a short introduction, all participants played the first scenario in order to familiarize themselves with the user interface and the interaction mode. After completing this scenario, they played either the second or third scenario – the distribution was randomized. As an external resource, all participants received a checklist including basic rules for handling complaints (Stauss & Seidel, 2010). After the gaming session, the participants filled out several post experiment questionnaires to secure their experiences and perceptions during the game. In addition, they were given the opportunity to describe their experience with and impression of the game in their own words.

Method of analysis. Like in the previous study, a mix of subjective and objective measures was used, which is particularly important in the analysis of serious games, as both the learners' feelings and motivation and objective criteria such as performance are relevant in this field (see Chapter 1.1.4). The subjective measures include the questionnaires (demographic data and prior experience, GEQ, EduTechRPG questionnaire, and Holtgraves questionnaire; detailed information about these questionnaires can be found in Chapter 1.1.4) and the qualitative interviews. The interview was mainly used to:

a) identify possible weak points, fix bugs, and improve the overall game experience afterwards,
b) find out which strategies and problem-solving approaches players apply when problems occur during the conversation and if these strategies are successful,
c) find out how the participants of the study perceived their own performance during the game.

Objective measures were mainly based on the evaluation of the chat protocols. They were primarily used to complement the results from the questionnaires and interviews and analyze the performance of the chatbots. The chat protocols were evaluated in regard to the answer quality of the chatbots. Three categories of answers were defined: *constructive*, *comprehensible*, and *nonsensical*. Each of the chatbot answers was assigned to one of these categories through human coding, which was based on a clear operational classification based on the works of Shawar and Atwell (Shawar & Atwell, 2007). Furthermore, the frequency of uses for each of the sentence openers was counted in order to find out which of them were used frequently, rarely, and not at all. The success in the game was operationalized in two ways: (1) the relative score (absolute score divided by the number of text inputs, and (2) the

[4] See Chapter 1.1.4 for detailed information about the questionnaires

total number of text messages because a rapid completion of the scenarios is assumed to indicate effective complaint management.

Evaluation results. Significant differences between the results of participants of the two groups could only partially be found. There were significant differences in the GEQ dimension *negative affect* but not in the dimensions *tension* and *challenge*. Furthermore, significant differences in the dimensions *thoughtful*, *polite*, *responsive*, and *engaging* of the Holtgraves questionnaire, but not in the dimension *comfortable*. The lack of more significant results could have been caused by two major issues: First, the number of participants was rather small (methodical condition). Second, the participants were asked to evaluate the perception of both scenarios played (either scenario 1 and 2 or scenario 1 and 3) combined (possible experimental design issue). Nonetheless, both groups perceived the chatbots as *human* and *skilled* irrespective of the conversational style and attitude, which indicates a high quality of the chatbot implementation. However, no significant results regarding the performance between experienced and inexperienced participants could be found, because the small sample size (n=4) did not allow for statistical tests. In general, although no generalization can be made due to the small sample size and some of the original assumptions could not or only partially confirmed, the results of the study are promising and showed that the approach of the game in general was assessed positively and the scenarios were perceived as realistic. It could be validated that the chatbot's discussion style has an influence on the players' perception of both the chatbot and the game experience, which underlines the successful dialog modeling.

1.2.3 Dimensions of the Design of Serious Role-Playing Games for the Training of Social Skills

Based on the experience gained during the creation of the framework, the development of the application scenarios, and the conduction of the evaluation studies, a set of dimensions or challenges that constitute important aspects in relation to the design of serious role-playing games for the training of social skills has been derived (detailed explanations on the single points can be found in Chapter 2) (Othlinghaus-Wulhorst & Hoppe, 2020):

1. *Learning context*: The learning context represents the basis of a serious game and relates to its theoretical foundation and the desired learning outcomes. Serious games have a great motivational potential, but this potential needs to be used to convey the pedagogical goals and learning objectives, i.e., the goal of the game should be aligned to learning outcomes as much as possible. The learning objectives need to be mapped onto concrete game mechanics, and learners should be given the opportunity to put the skills to be learned into practice in order to facilitate their acquisition (Naido, Ip, & Linser, 2000). Furthermore, it is important that the game's setting is appropriate for the learning context (Whitton & Hollins, 2008) and the selected topic (Salen & Zimmerman, 2004), and the chosen scenario and storyline need to fit the thematic context and use real concepts (Pivec, 2009). Moreover, the educational design must be based on an underlying corpus of background theories, otherwise the desired outcome cannot be ensured.

2. *Technical architecture and set-up*: The technical architecture refers to the concrete technical implementation of a serious role-playing game and the underlying system architecture, as well as the technologies and tools used for it. In general, major technical and implementational issues are flexibility, reusability, and extensibility/adaptability. In addition, the use of chatbots in such games entails some major challenges, such as natural language processing, for which different approaches and technologies are available.
3. *Dialog models and degrees of freedom*: This aspect relates to how the communication between the player(s) and non-playing virtual characters is carried out, structured, and controlled. In general, there are different interaction modes available (Brusk & Björk, 2009). In dialog-centric play settings, there is a range of communication models from fully pre-defined single choice inputs based on underlying conversation trees to sentence openers to free text input. Sentence openers seem to be a good compromise because they limit the possible inputs (making it easier to understand the general gist of a message) while still offering players a means to formulate their own inputs and express themselves more freely.
4. *Feedback and scaffolding elements and mechanisms*: These elements are essential for the transfer of learning to the real world and can be integrated in many different ways. In general, it can be differentiated between *ingame feedback* (implicit feedback during the role play session) and *aftergame feedback* (overall summary and specific annotations provided after the role play session). Both are important factors for enabling reflection. *Scaffolding* refers to intelligent mechanisms for support and guidance, which are provided to enable learners to master the challenges of the game and achieve the desired learning goals. Adaptation and personalization are key factors for education, which means that a serious game should adapt to the learners' level of knowledge, skills, progress, and performance (Bellotti, Berta, De Gloria, & Primavera, 2010).
5. *Relation between immersion and reflection*: The relation between these two components refers to the general question whether phases of immersion and reflection should overlap or occur separate from each other. While there are approaches claiming that it might be beneficial to have reflection processes take place within the game itself (Yusoff, Crowder, Gilbert, & Wills, 2009), there is reason to assume that immersion is actually hindering critical self-reflection (see Chapter 1.2.1.1). Accordingly, phases of immersion and reflection should be separated from each other and enforce role distance, allowing the learners to take over an outside perspective during the reflection phase.
6. *Collaboration support*: Collaboration support relates to the number of human players involved and the question whether it supports collaborative learning. There is a grown interest in the use of collaborative technologies for learning scenarios and the positive effects of collaborative learning have already been confirmed by recent research (Whitton & Hollins, 2008). In general, the integration of collaborative elements in a serious game may increase motivation and foster the development of cognitive skills (Romero, Usart, Ott, Earp, & de Freitas, 2012), and also provides multiple perspectives, create self-awareness of the learning process, thus make learning more authentic and

relevant (Whitton & Hollins, 2008). However, not in all scenarios it makes sense to include collaborative elements – it always depends on the context, the scenario and the learning objectives. There should always be a common goal and the tasks incorporated in the game should only be solvable by the players working together (Wendel, Gutjahr, Göbel, & Steinmetz, 2013), otherwise it does not make sense to integrate collaborative elements.

This set of dimensions and challenges in the design of serious role-playing games for the training of social skills constitute important aspects for the conceptualization, description and comparison of such games and can serve as a solid base for game developers and designers in this field. It is very important to deal with these aspects during the design process of serious role-playing games and to make informed decisions after contrasting different approaches and strategies and selecting those that are most suitable in the given context.

1.2.4 Contributions of the included publications to different research topics

The following compilation (Table 2) summarizes the main research topics the selected publications building the core of this thesis are focusing on, and maps them onto the dimensions described in the previous section. While all of them are located in the same subject area, follow the same conceptual and technical approach and generally deal with the same aspects, the concrete research objectives and fields of applications vary and also the particular focus is different between the publications.

The core of this thesis is made up of the following four publications:

[OH20] Othlinghaus-Wulhorst, J., Hoppe, H. U. (2020). A Technical and Conceptual Framework for Serious Role-Playing Games in the Area of Social Skills Training. Frontiers in Computer Science. 2(28), doi: 10.3389/fcomp.2020.00028.

[OJHH18] Othlinghaus-Wulhorst, J., Jedich, A., Hoppe, H. U., Harrer, A. (2018). Using Eye-Tracking to Analyze Collaboration in a Virtual Role Play Environment. In *International Conference on Collaboration and Technology* (pp. 185-197). Springer, Cham. doi: 10.1007/978-3-319-99504-5_15

[OH16] Othlinghaus, J., Hoppe, H. U. (2016). Supporting Group Reflection in a Virtual Role-Playing Environment. In *International Conference on Intelligent Technologies for Interactive Entertainment* (pp. 292-298). Springer, Cham. doi: 10.1007/978-3-319-49616-0_30

[OMH19] Othlinghaus-Wulhorst, J., Mainz, A., Hoppe, H.U. (2019). Training Customer Complaint Management in a Virtual Role-Playing Game: A User Study. In *European Conference on Technology Enhanced Learning* (pp. 436-449). Springer, Cham. doi: 10.1007/978-3-030-29736-7_33

It is important to mention that the publications are not embedded in this cumulative dissertation in chronological order; instead they are ordered and grouped based on their content: Although the journal article (Chapter 2) is chronologically the last one, it is the first to occur in this thesis as it provides a comprehensive overview of the developed framework including its conceptual and technical components as well as general dimensions and challenges in the design of serious role-playing games. Thus, it constitutes the basis for the

concrete scenarios and case studies presented in the subsequent chapters. Chapter 3 describes a case study based on a scenario designed for the training of conflict management skills in a workplace-oriented context. A first version of this application has been developed in 2012 (Emmerich, Neuwald, Othlinghaus, Ziebarth, & Hoppe, 2012), and after several re-iterations it has been evaluated within in the scope of the eye-tracking study in 2018 with focus on analyzing collaboration quality in this collaborative scenario. Chapters 4 and 5 both deal with a case study based on a scenario designed for the training of customer complaint management, but the focus is different. While Chapter 4 represents an earlier, more preliminary work describing the general characteristics of the original prototype and mainly focusing on the technically supported group reflection process, Chapter 5 focuses on the evaluation of playability, game experience, and perception of the virtual role-playing environment as well as the interaction with the different chatbots integrated in the scenario.

The other publications have not been included as dedicated chapters in this thesis, for different reasons: The article of Emmerich et al. (Emmerich, Neuwald, Othlinghaus, Ziebarth, & Hoppe, 2012) is relevant in the context of this work as it laid the cornerstone of this work, but it has been published before the actual start of the research series presented in the thesis. The work of Doberstein et al. (Doberstein, et al., 2016) is relevant in the context of this thesis too, but it is only a demo paper and the subsequent works (Othlinghaus & Hoppe, 2016) (Othlinghaus-Wulhorst, Mainz, & Hoppe, 2019) provide much more elaborate research works concerning the presented scenario. The article of Harbarth et al. (Harbarth, et al., 2018) is only peripherally related to this research work, namely in regards to mix-methods studies in online learning environments.

Publication / *Main Focus*	**Chapter 2:** [OH20]	**Chapter 3:** [OJHH18]	**Chapter 4:** [OH16]	**Chapter 5:** [OMH19]
Multi-agent Architecture *Technology/Implementation* (dimension 2)	X			X
Chatbot Implementation *Technology/Implementation* (dimension 2/3)	X			X
Automated Performance Analysis *Technology/Implementation* (dimension 2/4)	X			
Feedback Generation *Technology/Implementation* (dimension 2/4)	X			X
Group Reflection Support *Conceptual Approach* (dimension 5)			X	
Adaptive Feedback *Conceptual Approach* (dimension 4)	X			
Collaboration Support *Conceptual Approach* (dimension 6)		X		
Scenario Design *Conceptual Approach* (dimension 1)		X		X
Levels of Difficulty *Conceptual Approach* (dimension 4)				X
Game-based learning *Theoretical Background*	X			
Frameworks for Designing Serious Games *Theoretical Background*	X			
Social Skills Training *Theoretical Background* (dimension 1)	X			
Conflict Management *Theoretical Background* (dimension 1)		X		
Customer Complaint Management *Theoretical Background* (dimension 1)			X	X
Mixed-method Approach *Evaluation/Methodology*		X		X
Eye-tracking *Evaluation/Methodology*		X		
Collaboration Quality *Evaluation/Methodology*		X		
Collaboration Awareness *Evaluation/Methodology*		X		
Game Experience *Evaluation/Methodology*				X
Usability *Evaluation/Methodology*				X
Chatbot Perception and Performance *Evaluation/Methodology*		X		X

Table 2: Contributions of the included publications to different research topics

2 A TECHNICAL AND CONCEPTUAL FRAMEWORK FOR SERIOUS ROLE-PLAYING GAMES IN THE AREA OF SOCIAL SKILLS TRAINING

This paper was published in the journal *Frontiers of Computers Science* as part of an article collection on serious games (category: *Original Research*). This journal publishes research across all areas of fundamental and applied computational sciences, focusing especially on the application of computer science to other research domains to foster interdisciplinary research within computational sciences. *Frontiers* ranks as the 5th most-cited publisher among the 20 largest publishers in 2020. The paper was accepted on June 24, 2020 and published on July 31, 2020.

Othlinghaus-Wulhorst, J., Hoppe, H. U. (2020). A Technical and Conceptual Framework for Serious Role-Playing Games in the Area of Social Skills Training. Frontiers in Computer Science. 2(28), doi: 10.3389/fcomp.2020.00028 © 2020 Othlinghaus-Wulhorst and Hoppe.

Frontiers in Computer Science works under a Creative Commons License, Attribution 4.0 International (CC BY 4.0).

Author	**Contribution**	**%**
Julia Othlinghaus-Wulhorst	- Main responsibility in the design, implementation and evaluation of the games ColCoMa and CuCoMaG - Co-development of the framework - Writing of main manuscript	80%
H. Ulrich Hoppe	- Co-development of the framework - Research planning and guidance - Text revision	20%

frontiers in Computer Science

ORIGINAL RESEARCH
published: 31 July 2020
doi: 10.3389/fcomp.2020.00028

A Technical and Conceptual Framework for Serious Role-Playing Games in the Area of Social Skill Training

Julia Othlinghaus-Wulhorst and H. Ulrich Hoppe*

COLLIDE Group, Department of Computer Science and Applied Cognitive Science, Faculty of Engineering, University of Duisburg-Essen, Duisburg, Germany

OPEN ACCESS

Edited by:
Carlos Vaz De Carvalho, Polytechnic Institute of Porto, Portugal

Reviewed by:
Patricia Martin-Rodilla, University of A Coruña, Spain
António Fernando Coelho, University of Porto, Portugal
Ramon Fabregat Gesa, University of Girona, Spain

***Correspondence:**
Julia Othlinghaus-Wulhorst
othlinghaus@collide.info

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *28 December 2019*
Accepted: *24 June 2020*
Published: *31 July 2020*

Citation:
Othlinghaus-Wulhorst J and Hoppe HU (2020) A Technical and Conceptual Framework for Serious Role-Playing Games in the Area of Social Skill Training. Front. Comput. Sci. 2:28. doi: 10.3389/fcomp.2020.00028

Virtual role-playing games can provide an authentic experience of situated learning and allow for trying out different problem-solving and communication strategies without consequences in the real world. This is of particular interest and benefit for the training of social skills. This article presents a conceptual and technical framework for serious role-playing games for the training of specific social skills in virtual 2D learning environments involving chatbots in dialog-centric settings. It summarizes different use cases and evaluation results from prior studies. From the design perspective, several distinctive conceptual features characterize our framework: (1) chat-like interaction with an AI-controlled chatbot, (2) separate phases of immersion and reflection to facilitate a change of perspective that is considered conducive for learning, (3) the learning process is emphasized by means of adaptive feedback based on individual analyses. We propose a system architecture that is based on three components: (1) AI-controlled chatbots that adapt to the player's behavior, (2) a multi-agent blackboard system as the backbone in order to keep components independent and optimize performance due to parallel processing, and (3) intelligent support for an automated evaluation of the player's performance and feedback generation. The training scenarios presented and discussed in this article include workplace-oriented conflict management, patient-centered medical interviews, and customer complaint management. First evaluation studies indicate that the scenarios may be well-suited for real training situations. Due to its flexible architecture, our framework and approach can easily be tailored to different settings and use cases and thus serve as a basis for future research focusing on the adaptation to other contexts and systems. On the basis of these developments, we elaborate important design dimensions, reflect and discuss general issues and major challenges, summarize and contrast different approaches and strategies, and identify opportunities for serious role-playing games in the area of social skills training.

Keywords: virtual role play, intelligent support, serious games, social skills, chatbots

INTRODUCTION

In recent years, serious games have been established as an efficient medium in education and professional training (Michael and Chen, 2006; Marr, 2010). The serious gaming approach attempts to use the appeal of digital games not only for entertainment purposes but also to convey "serious" content and to train practice-oriented skills (Ritterfeld et al., 2009). The combination of the serious gaming approach with role play scenarios is particularly promising. Role play enables learners to explore new situations and train how to act and react in these situations (Martens et al., 2008). Virtual role-playing games provide mobile, safe, and continuable environments, whereas traditional role plays can be time-consuming, costly, and difficult to administer (Totty, 2005). In addition, they lack repeatability. One general problem in the evaluation of role play experiences for educational purposes is the effort involved in analyzing and reflecting on the actual role play following the enactment. Traditional scenarios typically rely on video recording and, if applicable, note-taking. However, virtual learning environments enable structured recordings with integrated indexing, navigation instruments, search functions, and cross-references between different media and data sources. In addition, computer-supported analyses can help to evaluate and track the learners' performance. This is an important aspect, since without feedback and post-role-play reflection, the transfer to real word situations cannot be ensured (Lim et al., 2009). An additional important advantage of serious role-playing games in contrast to other virtual learning activities and environments is the motivational component, which may lead to intense and passionate involvement of learners (Susi et al., 2007).

Based on a series of different instances of role-playing games for the training of specific social skills, this article presents the underlying conceptual and technical framework that facilitated the implementation of the different applications. This framework is characterized by using scripted chatbots as training cases in a dialogic setting. A multi-agent architecture supports both the actual dialogic processing as well as the evaluation of the dialogs and the generation of adaptive feedback. Conceptual and technical aspects of this framework are described in chapters Framework: Conceptual Approach and Framework: Technical Approach, following up on a discussion of related work in this area (chapter Related Work). Chapter Case Studies assembles several case studies conducted with different instances of virtual role-playing environments based on the framework, reporting on experience and evaluation results. Chapter Dimensions of the Design of Serious Role-Playing Games for the Training of Social Skills combines this specific experience with general issues in the design of serious role-playing games to devise a set design dimensions in the sense of important aspects to be considered in the design, description and comparison of serious role-playing games.

RELATED WORK

Serious Role-Playing Games for the Training of Social Skills

Serious games can be defined as "any form of interactive computer-based game software of one or multiple players to be used on any platform and hat has been developed with the intention to be more than entertainment" (Ritterfeld et al., 2009) and with an explicit focus on education. Games of this category are supposed to convey specific knowledge or train certain skills by using the attractiveness of entertainment games (Susi et al., 2007). Serious games can generally cover many different subject areas, but their application is mainly found in healthcare, education, and training, including military or employee training in companies (Marr, 2010). Serious games are widely accepted as an important and efficient medium with respect to education, training, and behavioral change (Michael and Chen, 2006). They are recognized to have several benefits: Serious games facilitate learning experiences while not having negative or harmful impacts (Ritterfeld et al., 2009). Games in general not only have a positive effect on the development of the player but can also be conducive to many different skills. Among others, Mitchell and Savill-Smith suggest that such target competences can be related to cognitive, social, analytical, and strategic aspects (Mitchell and Savill-Smith, 2005). Squire and Jenkins also made a comparable assessment (Squire and Jenkins, 2003). Further advantages include the reduction of costs and time associated with the use of serious games. They make it possible to recreate situations or working conditions that would otherwise not be possible in the real world (Corti, 2006; Susi et al., 2007). Serious games intend to facilitate deep and sustained learning (Gee, 2007) and prove to be more effective than traditional pedagogy and other educational technologies (Prensky, 2000; Ritterfeld et al., 2009).

Michael and Chen differentiate between games that educate and games that train (Michael and Chen, 2006). Games that educate should convey knowledge, facts and processes in a playful way, thereby contributing to education, while games that train are intended to improve the learners' skills in virtual environments or simulations. Our work is focused on the second category, more specifically on serious games for the training of social skills based on role play. Social skills can be seen as a sub-category of soft skills. The term *soft skills* refers to a broad concept that describes a set of personal attributes or traits expressing how persons know and manage both themselves and their relationship with other people (Dell'Aquila et al., 2017). While no universal definition of the term "soft skills" is available, Dell'Aquila et al. combine several different approaches to the following definition (Dell'Aquila et al., 2017): "Soft skills are not domain or practice specific; experientially based; both self and people orientated; goal-related behaviors; inextricably complementary to hard technical knowledge and skills enabling completion of activities and accomplishment of results; and crucial for effective leadership performance." Social skills refer

to soft skills related to interaction with other people. It describes "the ability to interact with others in a given social context in specific ways that are societally acceptable or valued and at the same time personally beneficial, mutually beneficial, or beneficial primarily to others" (Combs and Slaby, 1977) and includes, e.g., communication, cooperation, assertion, responsibility, empathy, engagement, and self-control (Gresham and Elliott, 2008). Role play is a great instrument to train interaction with other people. Assuming roles provides the opportunity to train to act and re-act in new situations. It facilitates the creation of knowledge and meaning through concrete experiences (Lim et al., 2009). Also, the observation of role play can lead to conclusions about own behavior (Martens et al., 2008). The integration of role play in a serious gaming context seems to be particularly promising, as this combination (a) incorporates a highly motivational character and (b) creates opportunities for exploration and experimentation in a protective environment without any consequences in the real world. In addition, virtual role plays may be much more effective than conventional approaches in settings where the social component is a crucial factor (Lim et al., 2009).

Several serious role-playing games for the training of social skills are available. They can be assigned to three main categories of relevant social skills: (1) leadership skills, (2) communication skills, and (3) conflict management. Examples for serious role-playing games for training leadership skills are *Virtual Leader* (Knode and Knode, 2011), *TeamUp* (Bezuijen, 2012), and *Learn to Lead* (Di Ferdinando et al., 2011). *Virtual Leader* is a simulation game in which students practice leadership styles and approaches within a 3D environment using avatars and intelligent agents in order to create a preferably realistic environment (Knode and Knode, 2011). Players participate in virtual business meetings with animated characters and are required to make a series of decisions in five scenarios with increasing complexity. *TeamUp* is a collaborative game for the training of teamwork and leadership skills, developed at the TU Delft (Bezuijen, 2012). In this game, four players need to work together to overcome several challenges, each designed to cover a specific element of effective teamwork. In *Learn to Lead*, the players have to lead a simulated team of employees (e.g., workers in a bank, a post-office, or a local government office) that is competing against other teams (Di Ferdinando et al., 2011). In this game, the players have two main objectives: First, they need to ensure that the company is running efficiently and productively. Second, they need to ensure that their teams develop in the desired manner. The Productive Leadership Game is a simulation game that is supposed to foster leadership competencies to improve team-based and organizational productivity (Kesti et al., 2017). A recapitulatory overview of serious role-playing games for training leadership skills can be seen in **Table 1**.

There are various examples for serious role-playing games aiming at the training of communication skills: *ENACT* (Marocco et al., 2015) is an online game for the standardized psychometric assessment and training of negotiation skills based on Rahim's model of conflict handling styles (Rahim and Bonoma, 1979). In this game, players assume different characters to negotiate with computer-controlled virtual 3D agents in various scenarios representing everyday life situations. They can always choose one of four possible pre-defined sentences to communicate with the agents. In *DREAD-ED*, players become part of a crisis management team that is dealing with an emergency situation (Haferkamp et al., 2011). The game is organized into a series of timed rounds, separated by phases in which a tutor can provide feedback to the players. Bosse et al. developed a game targeted at police academy students that focuses on decision-making aspects in critical situations like the so-called "door scene" in which a police officer has been informed about an incoming emergency call and is supposed to find out if it is indeed a case of domestic violence or not (Bosse and Gerritsen, 2016). The players interact with virtual characters in a realistic 3D environment by using a relatively simple interaction paradigm based on multiple choice and dialog trees. In the game *deLearyous*, players assume the role of a manager who just announced that the parking facilities of the company are no longer free and needs to deal with the reaction of an employee (Vaassen and Wauters, 2012) by using unconstrained written natural language input. The design of the virtual character representing the employee is based on a framework for interpersonal communication called Leary's Rose (Leary, 1957). *JUST-TALK* is a serious game to train law enforcement personnel for encounters with persons showing symptoms of serious mental illness (Hubal et al., 2003). The players interact with these computer-controlled characters using spoken natural language. They are supposed to look for indications of particular forms of mental illness so that they can adapt their approach in an appropriate way und thus defuse the situation. In *POINTER*, a game developed for interview training targeted at police officers, the players assume the role of a police officer interacting with a subject in the context of a police interview (Linssen et al., 2014). The subject here is a virtual agent who is not cooperating during the interview. The players' task is to interact with the subject in a way that makes it cooperate in order to gather information from them. *ELECT BiLAT* is a simulation game in which soldiers practice bilateral engagements within a cultural context (Lane and Hays, 2008). The recruits are supposed to conduct meetings and negotiations with local leaders. *Maritime City* is a game targeted at social workers. It aims at training the ability to read emotional states of persons and improving communication skills in verbal and non-verbal forms (Flynn et al., 2011). In this game, players are asked to investigate a disturbance at a house where a woman is living with her two children and need to investigate a range of approaches for each part of the scenario. *TARDIS* is a scenario-based serious game simulation platform that supports social training and coaching in the context of job interviews (Gebhard et al., 2018). It is specifically intended to be used by young people and job-inclusion associations to explore, practice, and improve their skills in a diverse range of possible interview situations by interacting with virtual agents acting as recruiters. *Communicate!* is a serious role-playing game designed to support practicing interpersonal communication between health care professionals such as doctors, pharmacists, or psychologists and a patient or client (Jeuring et al., 2015). In the scenarios included in the game, the players find themselves in a consultation with a virtual character during which they can choose between

TABLE 1 | Serious games for the training of leadership skills (overview).

Game	Author	Use of AI	Mode	Learning objective	Underlying framework/model/theory
Virtual Leader	Knode and Knode (2011)	Yes	Singleplayer	Leadership styles	–
TeamUp	Bezuijen (2012)	No	Multiplayer	Teamwork, leadership skills	–
Learn to Lead	Di Ferdinando et al. (2011)	No	Singleplayer	Leadership skills	Full-range theory
Productive Leadership Game	Kesti et al. (2017)	No	Singleplayer	Leadership competencies	Human capital production function

various options. They receive immediate feedback through the utterances and emotions of the conversational partner. The game *SALVE* (Augello et al., 2016) is using AI-controlled chatbots participating in medical consultations and is based on the Social Practice Theory (Schatzki, 1996). In contrast, Even et al. developed a serious game primarily targeting schizophrenia patients to support rehabilitation programs for social skills (Even et al., 2016). This approach is combining role play with problem-solving exercises on which remediation therapies rely. A recapitulatory overview of serious role-playing games for the training of communication skills can be seen in **Table 2**.

Conflict management is an important social skill that has been the subject of serious role-playing games in the past. *Choices and Voices,* for example, is an interactive simulation game for preventing violent extremism. In it, players explore and discuss issues and influences leading to tension and disruption in communities (Memarzia and Star, 2011). In this game, players face several moral dilemmas in which their decisions determine the outcome of the game (for themselves, their family, and their friends). This is supposed to show the significant consequences real life decisions can have. The storytelling game *Façade* asks players to resolve a conflict between a married couple. Through communication with the conflicted parties, they are to investigate the causes of their issues and provide counseling (Mateas and Stern, 2003). The emphasis here is on believable characters, natural language conversation, and a dynamic storyline. In *Office Brawl* the player assumes the role of a mediator, who is moderating a conflict between two parties in a workplace-oriented setting, using AI-controlled virtual characters (Glock et al., 2011). As a project manager in the game, the player needs to handle an argument between two members of a team. *FearNOT!* is a virtual drama for anti-bullying education targeted at children (Aylett et al., 2005). In this game, the bullying behavior of one of the characters is leading to dramatic episodes. The victim is seeking advice of the player who can interact with this character by using free text input. It is supposed to allow children to explore what happens in bullying situations in which they take responsibility for what happens to a victim without feeling victimized themselves. The game *LOITER* lets prospective police officers enact street interventions with loitering juveniles (Linssen et al., 2014) and aims to improve their social awareness. Here, players can experiment with different ways of interacting with the juveniles. *Self City* is a serious game developed for emotionally impaired adolescents, which is supposed to help them develop skills such as process-oriented thinking and conflict resolution (Van Dijk et al., 2008). In this game, players can walk around online in a virtual city. On their way to the cinema, they experience challenging social situations and learn how to deal with them. Players are accompanied by a daemon that provides advice in conflict situations and suggests alternative actions. The *Junior Detective Computer Game* has been developed as part of a multi-component social skills intervention for children with Asperger syndrome (Beaumont and Sofronoff, 2008). Here, players take the role of a trainee at the Detective Academy and are taught how to recognize complex emotions in computer-animated and human characters. They need to complete several missions, such as dealing with bullying, playing with others, and trying out new things. A recapitulatory overview of serious role-playing games for training conflict management can be seen in **Table 3**.

Frameworks for the Design of Serious Games

There is a number of existing models and frameworks for the general design of serious games, which describe fundamental components of such systems and support formal approaches to game design. A very general approach is the so-called *MDA (Mechanics—Dynamics—Aesthetics) framework* (Hunicke et al., 2004). It proposes three different perspectives for understanding and designing games: *Mechanics* refer to the actual implementation of the game. They describe its particular components (actions, behaviors and control mechanisms) at the level of data representation and algorithms. *Dynamics* relate to the overarching design goals and run-time behavior of the mechanics acting on player inputs and each other's output over time. *Aesthetics* refers to the resulting game experience. They describe the desirable emotional responses evoked in players, when interacting with the game system. Although the MDA framework is widely accepted and practically employed, it has weaknesses and limitations (Walk et al., 2017): It focuses too much on game mechanics, neglecting many design aspects of games, including an over-arching narrative. Therefore, it is not really suitable for all types of games, including particularly gamified content or any type of experience-oriented design.

Another approach toward serious game design is the *Four-Dimensional Framework* suggested by De Freitas and Oliver (2006). It postulates four main dimensions of learning processes to be considered in the design process of serious games: the *context* in which learning takes place (e.g., classroom-based or outdoors, access to equipment, technical support), the *learner specification* (e.g., learner profile, pathways, learning background), the *mode of representation* (e.g., level of fidelity,

TABLE 2 | Serious games for the training of communication skills (overview).

Game	Author	Use of AI	Mode	Learning objective	Underlying framework/model/theory
ENACT	Marocco et al. (2015)	No	Singleplayer	Negotiation skills	Model of conflict handling styles
DREAD-ED	Haferkamp et al. (2011)	No	Multiplayer	Disaster Communication	Theories of crisis and emergency risk management
The "Door Scene"	Bosse and Gerritsen (2016)	No	Singleplayer	Communication skills (police domain)	Education program of police academy students
deLearyous	Vaassen and Wauters (2012)	Yes	Singleplayer	Communication skills (workplace)	Interpersonal circumplex (Leary's Rose)
JUST-TALK	Hubal et al. (2003)	Yes	Singleplayer	Communication skills (law enforcement)	–
POINTER	Linssen et al. (2014)	Yes	Singleplayer	Communication skills (police domain)	Cognitive model for social interaction
ELECT BiLAT	Lane and Hays (2008)	Yes	Singleplayer	Cultural social conventions (military domain)	–
Maritime City	Flynn et al. (2011)	Yes	Singleplayer	Communication skills (social work domain)	–
TARDIS	Gebhard et al. (2018)	No	Singleplayer	Communication skills (job interview)	–
Communicate!	Jeuring et al. (2015)	No	Singleplayer	Communication skills (health care domain)	–
SALVE	Augello et al. (2016)	Yes	Singleplayer	Communication skills (healthcare domain)	Social practice model
Serious game for schizophrenia patients	Even et al. (2016)	No	Singleplayer	Communication skills (emotion recognition)	Social skills programs for schizophrenia

TABLE 3 | Serious games for the training of conflict management (overview).

Game	Author	Use of AI	Mode	Learning objective	Underlying framework/model/theory
Choices and Voices	Memarzia and Star (2011)	No	Singleplayer	Prevention of violent extremism	National curriculum
Façade	Mateas and Stern (2003)	Yes	Singleplayer	Conflict resolution	–
Office Brawl	Glock et al. (2011)	Yes	Singleplayer	Mediation	–
LOITER	Linssen et al. (2014)	Yes	Singleplayer	Street interventions	Cognitive model for social interaction, Virtual Storyteller (VST)
Self City	Van Dijk et al. (2008)	No	Singleplayer	Process-oriented thinking, conflict resolution	Dialogical self-theory
Junior Detective Computer Game	Beaumont and Sofronoff (2008)	No	Singleplayer	Bullying, conflict resolution	Social skills programs for individuals with Asperger syndrome

interactivity, and immersion used in the game), and *pedagogic considerations* (e.g., learning models, approaches for learning support). Like the MDA framework, this framework is a high-level model, meaning that it specifies a limited number of generic concepts that can or should be taken into consideration when designing or evaluating serious games, but only on a very general level with no concrete design or evaluation guidelines (Mayer et al., 2014).

This also applies to the *RETAIN (Relevance Embedding Translation Adaptation Immersion & Naturalization) model* by Gunter et al. (2006). This model was developed to support game development and to assess whether a serious game is appropriate for educational purposes, how well the academic or pedagogical content is immersed and embedded in the game's narrative and how knowledge transfer is promoted. *Relevance* means that the information students learn in the game should be relevant to the game world as well as to the players' targeted objectives. *Embedding* should be done in a way that learning objectives and fantasy are tightly coupled. *Transfer* refers to how well players can recognize and apply newly learned information outside the game environment. *Adaptation* means that players apply their learned knowledge to create new scenarios that apply literacy skills in a new domain. *Immersion* should be facilitated by the game environment and the ability to create customizable social presence. *Naturalization* means that players should be encouraged to gradually use their own skills to gain the

knowledge necessary for success in other problems and subject areas (Kenny and Gunter, 2011).

The *Triadic Game Evaluation* (TGE) (Harteveld, 2011) approach stresses three different perspectives for the design and evaluation of serious games: reality, meaning and play. The *reality* component determines the game subject, variables and definitions. It could be represented by players from the real world or a representation of the real world inside the game. Evaluation criteria in regards to this component include fidelity, realism, and validity. The *meaning* component of the framework considers how a meaningful effect beyond the game experience can be achieved and incorporates aspects such as communication, learning, rhetoric, and opinions. Evaluation criteria include reflection, transfer, and relevance. The *play* component refers to the fact that games are primarily highly interactive and engaging tools that immerse players into a fictitious situation, and is related to game elements like actors, rules, resources, challenges, and competition. Evaluation criteria for this component are engagement, fun, and immersion. The TGE framework claims that games need to be designed equally along these three components (Kortmann and Harteveld, 2009). In contrast to the aforementioned models, this framework comes with a concrete agile development model that describes different software engineering phases and decision moments in the creation process. However, specific design and implementation guidelines are not included.

In summary, the various promising approaches to training social skills by means of role-playing games are still defined on a very general level. Our aim is to provide a comprehensive conceptual and technical framework for the concrete design and implementation of serious role-playing games for the training of social skills in dialog-centric settings with virtual characters through which we would support more efficient and effective design and implementation of such game environments.

FRAMEWORK: CONCEPTUAL APPROACH

From the design perspective, several distinctive conceptual features characterize our framework: (1) chat-like interaction with an AI-controlled chatbot, (2) separate phases of immersion (role-playing) and reflection to facilitate a change of perspective that is considered conducive for learning, (3) the learning process is emphasized by means of adaptive feedback based on individual analyses.

Chatbots in Virtual Role-Playing Environments

Chatbots are computer programs (conversational agents) that communicate with users in natural language. Their purpose is to simulate a human conversation via text or voice interactions. Originally, chatbots were developed for entertainment purposes. However, especially in today's world, in which the possibilities of computer use are becoming more and more diverse, the use of chatbots can be extended to many other areas. Chatbots are found in daily life now, such as personal assistants (like Google Assistant, Amazon Alexa, or Apple's Siri), search engines, customer service and support, and healthcare coaching (Winkler and Söllner, 2018). They can be used in a variety of domains including business, e-commerce, entertainment, medicine, and others (Kerly et al., 2006; Shawar and Atwell, 2007).

Chatbots can also be used successfully for learning. Past studies even show that chatbots present feasible means to improve learners' results (Kerly et al., 2006). They have been used for a variety of purposes including medical education and therapy, language learning, as well as receiving feedback and strengthen motivation and self-efficacy (Winkler and Söllner, 2018). Chatbots have also been used in serious role-playing games, as shown in the examples in chapter Related Work. The use of chatbots in serious role-playing games has several advantages. First, having a chatbot interact with the player instead of a human ensures a certain level of standardization that could never be achieved in a setting with human actors. Second, scenarios including a chatbot are repeatable, independent of time and place, and no additional resources are needed. An important part of chatbots is the creation of dialogs. A chatbot can only be as good as its knowledge base used for answer generation (Abdul-Kader and Woods, 2015). The problem of the "classic" chatbots is that they do not allow to store the course of the conversation and have no real understanding of the answers. However, a realistic and responsive behavior of chatbots is important to increase the players' engagement and contribute to the immersive nature of role plays. To achieve this, our approach proposes several technical workarounds that will be explained in detail in chapter Multi-Agent Architecture.

Immersion and Reflection

The educational impact of serious role-playing games highly draws on the "willing suspension of disbelief" by the players who commit to the role they are supposed to play (Lim et al., 2009). Thus, this kind of system intends to create a certain degree if immersion. Janet H. Murray defines the term immersion as follows: "A stirring narrative in any medium can be experienced as a virtual reality because our brains are programmed to tune into stories with an intensity that can obliterate the world around us… The experience of being transported to an elaborately simulated place is pleasurable in itself, regardless of the fantasy content. Immersion is a metaphorical term derived from the physical experience of being submerged in water. We seek the same feeling from a psychologically immersive experience that we do from a plunge in the ocean or swimming pool: the sensation of being surrounded by a completely other reality, as different as water is from air, that takes over all of our attention, our whole perceptual apparatus" (Murray, 2017). When players identify themselves with the character they are assuming in the game and are immersed, their motivation to proceed and succeed in the game increases (Annetta, 2010). This intrinsic way of motivating learners is something conventional instruction modes do not have (Yee, 2006). Players become immersed in a game because they find it satisfying, and through this intrinsic motivation, they get more engaged in the learning task (Annetta, 2010).

In terms of experience-based, authentic learning, it seems reasonable to carry out the enactment in an immersive situation. However, there is reason to believe that the immersion tends

to impede the critical self-reflection that is important for the learning process (Malzahn et al., 2010). Reflection is a successful tool to improve the learning process (Jonassen et al., 1993), and it is needed to ensure the transfer to real-life situations (Lim et al., 2009). During the reflection process, people recapture, rethink, and evaluate their experiences to develop new understandings and appreciations (Boud et al., 1985). It is to be expected that the amount of reactive attention required for immersion impedes the players' ability to distance themselves from the role, which in turn interferes with self-reflection. Thus, the requirement of role distance in phases of reflection suggests that the mode should be changed to help the learner step out of his role and adopt a different perspective. Based on this assumption, we decided to separate the actual role-playing game from the reflection session in our framework.

Adaptive Feedback

As stated above, an important challenge for serious role-playing games is shaping the narrative experience and the pedagogical outcomes that generally depend on post-role-play reflection and feedback (Lim et al., 2009). Feedback on the performance of the player(s) during the role-playing session is necessary to ensure the transfer to real-life settings. It is supposed to help learners to improve their performance by providing information about the correctness of their actions (Shute, 2008). Johnson et al. identified four feedback characteristics: (1) the *type* of feedback (e.g., outcome-based or process-based feedback), (2) the *timing* of feedback after an action (i.e., immediate or delayed feedback), (3) the *modality* in which the feedback is presented (e.g., spoken or text-based feedback), and (4) *adaptation* to learner characteristics (e.g., in regards to prior knowledge or spatial ability) (Johnson et al., 2017).

Our framework relies on adaptive feedback based on an automated, individual performance analysis. We differentiate between three types of feedback: The first one is implicit feedback during the role-playing session through the reactions of the chatbot (*ingame feedback*). These reactions can be non-verbal (e.g., facial expressions) or verbal. Real-life situations are simulated through both types of reactions to the players' actions. The second one is a general summary of the analysis results (*aftergame feedback*). Players should receive an overall feedback on their performance during the role play that summarizes the most important aspects (positive and negative). The third type is direct and specific feedback on single incidents during the role play that can be provided through prompts in a replay of the conversation. A replay offers several advantages: The whole conversation can be shown again step by step and augmented with individual feedback at certain points, commenting on specific actions of the player. Also, it provides the possibility to navigate between the different phases of a conversation, pause the replay, or jump to the next feedback marker. As a result, it is much more flexible and searchable than, e.g., a video of a conventional role play.

FRAMEWORK: TECHNICAL APPROACH

In our approach, the technical implementation of such systems entails three main challenges: (1) dialog modeling of the chatbot, (2) implementing a multi-agent system as the backbone in order to keep components independent and optimize performance, and (3) performance analysis and feedback generation. The following section will present our approach toward each of these aspects in detail.

Dialog Modeling

In our framework, the *Artificial Intelligence Markup Language* (AIML) is used for the implementation of the chatbots' conversational logic. It is a common XML-based solution for passive AI-controlled chatbots, which comes with an easy syntax and a small number of control structures (Wallace, 2004). AIML relies on a simple pattern matching. It consists of categories, each containing a pattern and a template. If the user input is matching a pattern, the template defines the answer or action to be given. Recursion and wildcards allow for many different inputs matching one single pattern, while the ability to store a context and the use of variables and conditions allow a complex and sophisticated chatbot design.

Although AIML has a long history and is a common solution for chatbots used in educational contexts, it has certain limitations. One problem is the passive nature of AIML. An AIML chatbot only reacts to an input it receives, it cannot take the initiative. This behavior can be bypassed by using external triggers to make the bot become active when required in certain situations. Another problem is that an AIML chatbot (as is true for all artificial natural language processing) cannot truly grasp the sense of what has been said. The AIML chatbot only checks the user input against predefined patterns; if there is no match, it can at most output some default statements (which need to be predefined as well). To solve this problem, our framework proposes the use of sentence openers in dialog-centric role play scenarios. This means that players always have to select a sentence opener from a predefined set and supplement it with free text input to compose a message.

This approach has several advantages: First, a sentence opener already defines the general gist of a message (e.g., affirmation, rejection, proposal, inquiry). As a result, it is at least possible to provide a default answer that is tailored to the selected sentence opener even if the free text input following the opener does not match a predefined pattern. Furthermore, if each phase of the chat conversation has unique sentence openers, the chatbot always has some kind of context information. Second, the use of sentence openers reduces the complexity of the dialog scripts dramatically because the possible starting points of all input sentences are already known. Third, sentence openers provide support to the players and help them phrase their messages. In addition, sentence openers improve the overall atmosphere of the simulated conversation and make it seem more realistic and natural. Last, sentence openers (in contrast to fully predefined text messages) still allow for free text input that can be analyzed in detail and influence the course of the game.

Multi-Agent Architecture

Our technical framework is based on a uniform multi-agent system architecture with a blackboard as the communication and integration mechanism. The blackboard is realized through a so-called *tuple space*. The components (agents) in this system are loosely coupled, i.e., they do not communicate with each other directly but only via entries on a central tuple space server (Gelernter, 1985). These entries have a simple tuple structure that contains primitive data types (integers, characters, booleans) and strings. According to the original concept of Gelernter, there are only a few generic operations (read, write, take, wait-to-take, etc.) to interact with such a blackboard. In contrast to a pure database solution, however, there are active trigger mechanisms such as notifications. The *SQLSpaces* developed in the COLLIDE group itself serve as a specific implementation basis in our framework (Weinbrenner, 2012). While the server itself is implemented in Java, the system framework of SQLSpaces provides clients for the agent programming in various programming languages. SQLSpaces also facilitates the logging of relevant data of each gaming session, which can later be used for analysis and comparison.

The overall system consists of a user interface and various agents, each of which is responsible for one task in either dialog analysis, feedback creation, or game control. The user interface in the three implemented training scenarios described in this article have been implemented as a web application using HTML, CSS, and JavaScript (2D frontend). Previous implementations were based on OpenSimulator3 (3D frontend), but since there were no specific advantages of 3D environments over 2D environments, we decided to go ahead with a 2D approach (Malzahn et al., 2010). As described above, the client (user interface) and all agents are writing and reading tuples from the tuple space server without communicating with each other directly, which results in a loosely coupled and adaptive system. That means, agents can easily be adapted, added, or removed depending on the actual application scenario.

The agents can be divided into three groups, depending on their functionality. *Pervasive agents* are overarching agents, which are crucial in connecting the individual game components. The *register agent*, for example, is managing the log-in of the player (or players in a collaborative scenario). When a new client is logging in, the register agent receives a request via the tuple space (callback) and starts a new gaming session. The *silence agent* reacts if a player has been inactive for a certain amount of time, in which case the agent is triggered and sends an internal message to which the chat bot responds. After the fourth internal message from the silence agent, the conversation ends. *Pre-processing agents* are used to pre-process the player's input before the answer to it is generated in order to provide the best possible answer. This pre-processing is mainly used to overcome the limited capabilities of AIML: Analyzing certain aspects separately helps to prioritize specific behaviors, i.e., make sure that the chatbot is reacting adequately to rude or aggressive behavior. In addition, this procedure reduces the structure of the AIML scripts and supports the feedback creation. Each of the implemented scenarios uses different pre-processing agents depending on the context. All pre-processing agents analyze the player's input regarding one specific aspect. **Figure 1** shows the basic architecture.

Performance Analysis and Feedback Generation

Both performance analysis and feedback generation always depend on the context and the learning objectives of the serious role-playing game. As described above, our architecture is using analysis agents, each of which is responsible for the evaluation of one specific aspect of the player's communication behavior. They are divided into pre-processing agents and regular agents.

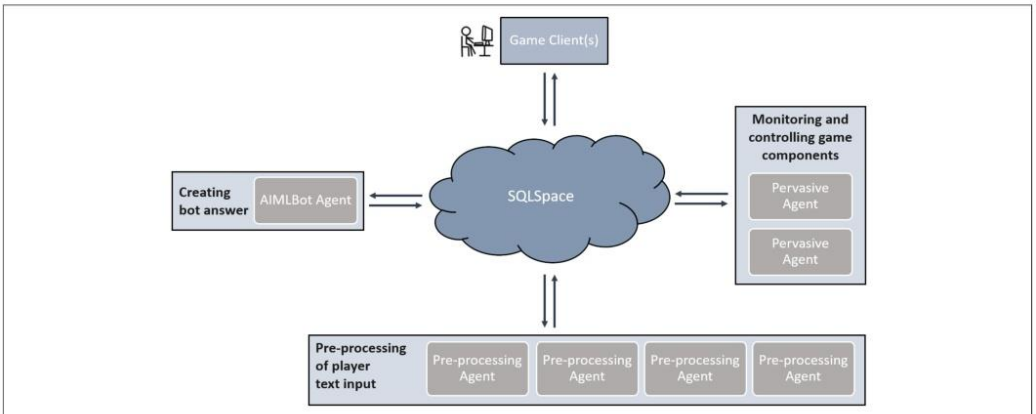

FIGURE 1 | Basic system architecture.

Pre-processing is necessary for generating a suitable chatbot response. For example, if a player acts aggressive or rude, the chatbot should react to this behavior regardless of any other information the player's message contains. The results of the pre-processing are collected, and if an immediate reaction to a specific behavior is required, the text input is modified. If, for instance, a swearword has been detected in a player's message, the complete input string is replaced by a specific trigger ("swearword"), causing the bot to react appropriately. The same applies to other behaviors. In case the pre-processing agents do not find anything that needs an immediate reaction of the chatbot, the bot receives the original text input. Simultaneously, all other analysis agents evaluate the message and add their feedback to it in the form of feedback tags (e.g., #praise#, #interruption#, or #criticize#). These feedback tags mark any situations in which the player is supposed to receive feedback during the replay that is taking place in the reflection phase following the role play session. The tags are filtered out during the chat session; the players do not get to see them during the game, but they play an important role in the feedback generation.

CASE STUDIES

Based on the framework described above, the research group COLLIDE at the University of Duisburg-Essen has conducted various case studies with different instances of virtual role-play environments. The training scenarios include workplace-oriented conflict management, patient-centered medical interviews, and customer complaint management.

Case Study: Conflict Management

ColCoMa (Collaborative Conflict Management) is a collaborative serious game for the training of conflict management strategies in an organizational context within a role-playing scenario, developed at the COLLIDE group in 2012. It involves two players in a conversation with an AI-controlled chatbot acting as a mediator in a 2D virtual environment. The following description of the approach and game design is based on the work of Emmerich et al. (2012).

Approach

In ColCoMa, two players have a conversation about a fictitious conflict, moderated by an AIML chatbot in the role of a mediator. The main goal of the players is to resolve the conflict by showing constructive and appropriate behavior during the conversation. Each player is assigned a predefined role in this fictitious scenario: As a member of the computer support hotline of a big software company, Mr. Meier is conscientiously taking much time for his customers. Mrs. Schmidt is his supervisor. She is dissatisfied with Mr. Meier's way of working. She notices that he takes too much time for the customers and therefore does not work efficiently in her eyes. Mr. Meier does not agree with her, and the situation escalates after a negative appraisal of Mr. Meier's performance. In order to support immediate understanding of the situation and empathy with the assigned role, the scenario is kept as simple and comprehensible as possible and focuses on the main conflict as well as the person's feelings.

Game Design

The players are introduced to the game and the scenario through a cartoon-like picture story that is told from their respective role's perspective and is supposed to result in conflicting points of view. The conversation itself takes place in a chat window where graphical representations of the mediator and the other player's character are shown to create the association of sitting opposite each other. The dialog partners can communicate via simple text messages. Facial animations can be evoked via common emoticons. The interface also includes a notepad with hints as well as a help section that offers additional information on the game controls and the fictitious scenario if needed. **Figure 2** is showing the basic user interface.

The conversation is divided into five phases according to Proksch (2010): (1) framing phase, (2) topic collection, (3) working on the conflict, (4) looking for a solution, and (5) contract. The framing phase represents the starting point of the mediation talk and is important for establishing certain rules for the conflicted parties and their behavior toward each other. The actual conflict is not yet the focus. Instead, the participants state their personal hopes and mediation goals and reflect on their own point of view as well as the opponent's position. In the second phase, both parties are supposed to name relevant topics they would like to put on the agenda during the mediation talk, like performance review, working conditions, the participants' perspective in the company as well as their behavior toward each other. The mediator chatbot recognizes the topics based on a list of keywords and phrases. In order to be able to advance in the game, the two players need to name three topics; otherwise the mediator terminates the conversation due to a lack of contribution. If only two topics are volunteered, the mediator will suggest a third one. The mediation talk itself takes place in the third phase. The main task during this phase is to discuss the selected topics in detail. Both players are given the opportunity to explain why a topic is important to them, what changes they would like to see in regard to the specific topic, and what they themselves can contribute to realize these changes. They are also given the opportunity to comment on whether the other party's perception is correct and to rectify their position if this is not the case. The aim of the fourth phase is to find solutions for the different topics that are acceptable for both parties. Finally, in phase five, they are supposed to agree to adhere to the solutions they came up with and enter into a contract.

The mediation talk is followed by a reflection phase in which both players receive feedback on their performance in order to help them reflect on their behavior. At the start of this phase, players get the opportunity to directly exchange feedback with each other in a free chat without the mediator. After this free chat, each of them receives an overall feedback on the own performance during the mediation talk. Finally, the players take part in a replay session of the whole chat conversation, but this time augmented with individual feedback commenting on especially positive and negative contributions of the players. A change of the graphical interface during the replay is supposed to reinforce role distance, which is assumed to be conducive for learning (see chapter Immersion and Reflection).

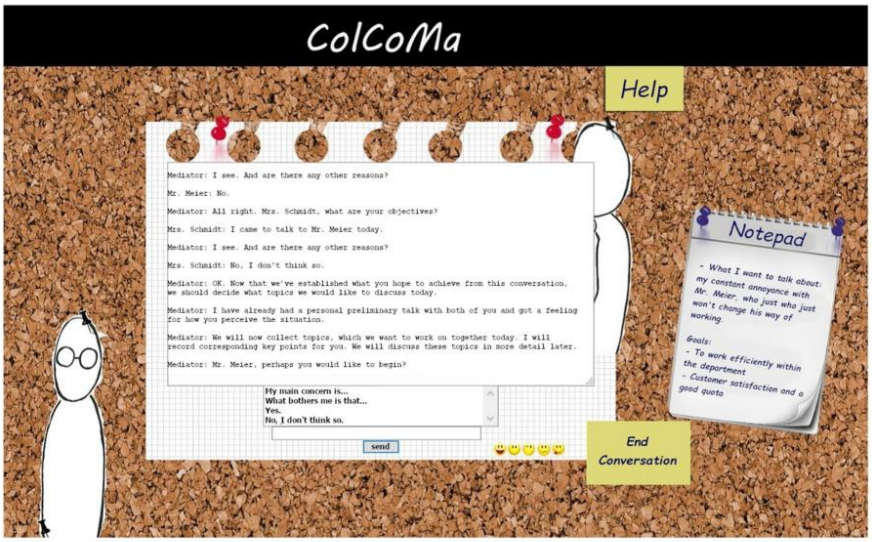

FIGURE 2 | ColCoMa chat interface.

The performance analysis and assessment is based on general rules that conflicting parties have to adhere to during a mediation talk, such as not being aggressive or rude, not being reproachful, and not impairing the opponent's autonomy (Stauss and Seidel, 2010). Instead, the participants are supposed to have an open and constructive attitude, name topics and issues in a concrete way, and help the other party understand their perspective. The evaluation of the players' performance during the mediation talk is done by several analysis agents, each responsible for one specific behavioral aspect, e.g., rudeness (by comparing the players' input to a list of swearwords and defamations), aggression (e.g., by checking for multiple exclamation marks or use of all-caps spelling), emotion-showing (e.g., use of emoticons), or the use of I- and you-statements (by counting the amount of words referring to the speaker and those referring to the dialog partner). Some of the analysis results are used just for the overall feedback that is provided to players after the conversation.

Evaluation Results

In 2018, an eye-tracking study has been conducted in collaboration with the Dortmund University of Applied Sciences and Arts (Othlinghaus-Wulhorst et al., 2018). The results of this study will be summarized and discussed in this section.

Apart from getting feedback on the prototype, the main goal of the study was to investigate the question if there is a correspondence between gaze synchronicity of the two players and the quality of collaboration. Twenty subjects (average 22.8, SD = 2.84, 5 females, 15 males) participated in the study and have been tested in dyads, using two desktop-based eye-trackers to track the players' gaze during the experiment. To investigate the research question, three main hypotheses have been examined: The first hypotheses postulated "a positive relation between the convergence of visual foci of attention (gaze synchronicity) and the successful completion of the game (achievement score)" (Othlinghaus-Wulhorst et al., 2018). In this study, *gaze synchronicity* has been defined as the extent to what the two players have been looking at the same areas of interest in the same time interval during the course of a gaming session. The so-called *achievement score* has been used to measure the success in the game and reflects the players' performance during the mediation talk based on three criteria: (1) automated feedback generated by the system, which summarizes the players' behavior during the game, (2) the successful completion of the topic collection phase (which has been considered a major milestone in the game), and (3) the successful completion of the game, which is achieved when both players sign a contract, which includes the agreements and rules they worked out together with the mediator. Referring to the hypothesis, a highly significant correlation between the gaze synchronicity and the achievement score has been found on the aggregate level (taking overall eye-tracking convergence as a global parameter).

In the second hypothesis it is assumed that "there is a positive relation between the convergence of visual foci of attention (gaze synchronicity) and the quality of collaboration in the chat." (Othlinghaus-Wulhorst et al., 2018). In order to define the *quality of collaboration*, a rating scheme has been developed, which includes five dimensions: (1) argumentation

(players discuss or bring forward justifying arguments), (2) agreement/disagreement (players endorse or dissent from one another), (3) collaborative orientation (players refer to each other, ask questions, provide feedback or refer to topics brought up by the other player), (4) solution orientation (players try to find or propose a solution), and (5) shared awareness/reinforcing shared history (players share common knowledge or explain their situation). Based on this scheme, all chat messages have been analyzed and checked against the five dimensions and assigned a total quality score. Finally, all matches of a gaming session have been added up to a percentage indicating the overall quality of the collaboration for a pair of players. Relating to the hypothesis, a high correlation between the gaze synchronicity and the collaboration quality has been found, especially for the dimension's agreement/disagreement, solution orientation and shared awareness.

The third hypothesis proposes "a dynamic (time-related) congruence between similar eye movements (synchronicity) and the quality of collaboration in the chat" (Othlinghaus-Wulhorst et al., 2018), meaning that there is not only a gaze synchronicity on the aggregate level, but also synchronicity between convergent eye-tracking and chat interaction during the course of the game. This hypothesis could not be verified. It is assumed, that the specific nature of the chat might be a reason for this, as three persons are involved (the two players and the mediator chatbot) and thus the two human actors do not really communicate directly, but only to the mediator. They answer his questions and do not really have the chance to communicate with each other directly, which is resulting in a predefined structure of the chat conversation and rather long time interval between the utterances of the two players.

Case Study: Patient-Centered Medical Interview

In 2013, a training scenario for medical interviews has been developed at the COLLIDE group. It is supposed to give medical students the opportunity to train doctor-patient conversations autonomously and systematically in the form of role plays with simulated patients. The following description of the approach and game design is based on the work of Behler et al. (2013):

Approach

This scenario for the training of doctor-patient communication has the basic goal to train the communication strategies between doctor and patient and is tailored to the target group of medical students. Here, the player takes the role of a locum doctor for family medicine whose goal is to uncover all the symptoms of a patient in a given time. To achieve this, they have to use methods of the GOG (*Gesundheitsortientierte Gesprächsführung*, engl. "health-oriented negotiation") (Schwantes and Kampmann, 2007), in order to create a pleasant conversation atmosphere. To successfully master the game, it is necessary to behave in accordance with this concept and to bring in the guidelines in the course of the conversation. Another important learning objective is to build trust and empathize with the patient, as these aspects play a central role in the doctor-patient conversation (Kruse, 2000). Medical diagnosis is not a learning objective in this scenario, so the game can be used independently of progress in medical studies.

Game Design

At the beginning, the player enters the waiting room, where several patients are already sitting and waiting for their call. The patients represent different scenarios, which differ in the content and level of difficulty. The level of difficulty is determined by the number of symptoms to be identified and the willingness to talk about his or her condition. In the waiting room, the trophies and high scores already achieved by the current player are also displayed. By choosing a patient, the player starts the scenario and enters the doctor's office, where the actual interview with the patient takes place.

In the office, the player communicates with the patient via text input. The player chooses a suitable sentence opener and completes the sentence with free text. The sentence openers are related to GOG phases. In addition to verbal interaction, the player can also use items from the doctor's bag (information leaflets, stethoscope, pills, and syringe) and conduct non-verbal actions like nodding, smiling or touching the patient, which are also important in real interpersonal communication (Ziebarth et al., 2014). The items provide a playful added value. The player has to find out when which item is reasonable to use and receives bonus points for this, but only in combination with appropriate topics—otherwise, points are deducted. **Figure 3** is showing the basic user interface of this scenario.

In order to win the game, i.e., to achieve the highest possible score, the player has to collect points for recognized symptoms as well as points for trust-building and empathic contributions and actions. The accumulated sum of trust and empathy points in the game represents the conversation atmosphere and serves as a threshold value that defines how quickly a symptom is revealed by the patient. The patient reveals symptoms when the player addresses a scenario relevant topic and has reached the corresponding threshold value.

The main conflict is between the limited time available to the player to find the symptoms and the patient who only reveals them under certain conditions. This situation resembles a doctor's real conflict between time pressure and the desire to help patients comprehensively. Each scenario of this game contains a side mission to increase replayability. While the main task includes finding relevant symptoms, a secondary task could be, e.g., to point out the benefits of assisted living to an elderly patient to ensure long-term care. Side missions give more depth to the game as they refer to the social situation of the patient and thus lead to more immersion (McMahon and Ojeda, 2008). Players receive bonus points and trophies for solving side missions. As an additional incentive system, the total number of points is entered to a leaderboard, which all players can see. According to Festinger's theory of social comparison (Festinger, 1954), this motivates players to improve their own abilities, which are represented by the points.

The gaming session is followed by a reflection phase. First, the players are presented with their individual score in the fields trust, empathy and symptoms. Afterwards, they receive detailed feedback in the form of an augmented replay, in

FIGURE 3 | The doctor's office (basic game interface).

which the analysis results are presented. In this analysis, the player interaction is, e.g., checked for the use of paraphrases, emotions expressed to the patient, showing choices to the patient, addressing him or her by name and the use of all phases of the GOG. Although the phases do not have to be passed linearly, goal guidance and explanations for example are particularly relevant toward the end of the conversation. Pauses, nods and facial expressions are evaluated as well. In addition, behaviors are taken into account that do not directly lead to an improvement of the score but influence the course of the conversation. For example, the patient reacts verbally to excessive talking of the player and a lack of balance between the doctor and the patient as the subjects of the player's statements. This leads to a loss of time, which increases the central conflict of the game. As in a real situation, the player receives his feedback directly from the patient and can react to it in the process of the conversation.

Evaluation Results

The prototype has been evaluated in two studies, both performed by the COLLIDE group in cooperation with the Department of Family Medicine of the Charité in Berlin (Ziebarth et al., 2014). The results of these studies will be summarized and discussed hereinafter:

The focus of the studies was the examination of usability and playability, as well as immersion and reflection. The following key questions have been deducted from the global objectives: (1) Does the flow of the game feel natural? (playability), (2) Is the player able to manage the game well? (usability), (3) Is the game immersive? (immersion), (4) Is the reflection phase at the end of the game perceived as helpful? (reflection), (5) Which functions are used?, (6) What are the difficulties in using them?, (7) How is the game perceived by the target group? *Playability* has been evaluated using self-created items relating to the clarity of the goal, structural problems regarding game flow (i.e., the use of sentence openers to create chat messages), "functional" playability (i.e., the extend of feeling understood by the patient), and the complexity of the game (Ziebarth et al., 2014). *Usability* has been assessed based on the following categories of ISO 9241-110: conformity with user expectations, suitability for learning, self-descriptiveness, and error tolerance. The items of the questionnaire were phrased based on the German inventories Isonorm[1] and IsoMetrics[2]. The aspect of *immersion* was measured based on the approach developed by Jennett et al. for measuring immersion in digital games (Jennett et al., 2008). The items selected for the studies address the subjective enjoyment of the game's representation, fun factor, immersion, and emotional involvement (Ziebarth et al., 2014). For the assessment of the *reflection* support, participants were asked what they thought the game is aiming to train, if they viewed the annotated replay, and if they thought about what they could have done differently (Ziebarth et al., 2014).

The first experiment was an observational study with 7 medical students (average 21, SD = 2.582, 6 females, 1 male). Although the results indicate that the idea and approach of the game were assessed quite positively, the observations showed slight problems with the general usage of the game. While the interaction principles have been generally well-understood, a few participants reported problems with expressing themselves using the predefined sentence openers. Also, the free text supplementing the sentence opener was often not understand by the chatbot, because some topics have not been considered in the

[1]https://abeto-online.de/ep/index,id,3314.html
[2]http://www.isometrics.uni-osnabrueck.de/

design. Apart from these limitations, the students liked the game as an alternative for the training of medical interviews before performing them with human patients. The second experiment was an online study with 21 medical students (average 23.05, SD = 4.295, 15 females, 6 males). The online questionnaires (n = 21) as well as the questionnaires completed by the participants on paper during the observation (n = 7) were included in the evaluation of the questionnaire in the subject areas playability, usability, immersion and reflection. The results mostly support the findings gained in the observation study. While most of the participants liked the user interface [M = 4.18 (of 5), SD = 0.819] and understood the goal of the game [M = 3.68 (of 5), SD = 0.905], its suitability for learning [M = 3.82 (of 5), SD = 0.782] and the self-descriptiveness [M = 3.33 (of 5), SD = 0.603] were considered good (above average), and the imaginative immersion [M = 3.29 (of 5), SD = 0.076] as well as the emotional involvement (M = 2.93, SD = 1.086) showed only average values. In addition to the questionnaires, a total of 36 conversation transcripts were evaluated in order to uncover possible weak points in the text recognition module of the system, which was used to fix and further improve the AIML scripts.

Case Study: Customer Complaint Management

The case study CuCoMaG (*Customer Complaint Management Group reflection*) is a serious role-playing game for the training of customer complaint handling based on theories of consumer psychology and complaint management, originally developed in 2016 at the COLLIDE group in the context of a student master project (Doberstein et al., 2016; Othlinghaus and Hoppe, 2016). It has been re-designed and evaluated in 2019 by Othlinghaus-Wulhorst et al. (2019). The following description of the approach and game design is based on these works.

Approach

In this game, the player assumes the role of a customer service employee in the fictitious company *LittleOnes*, a producer and seller of personalized clothing for children via an online shop. The player is confronted with a chatbot in the role of a complaining customer, who is reporting a certain problem. The player is communicating with the customer through a simple chat environment. Like in the other scenarios, the player has to select a sentence opener and supplement it with free text in order to formulate a chat message. The chat setting is ideal in this use case, as it simulates everyday work situations for people working in the customer support sector. **Figure 4** is showing the general user interface of this scenario. This game has one distinctive feature that sets it apart from the previous ones presented: It offers explicit support for group reflection (Othlinghaus and Hoppe, 2016). Group reflection enables a collective exchange and thus collaborative learning (Schuster, 2010). The group reflection session, which is supposed to take place subsequent to the role play session, is designed to be guided by a trainer. This trainer is given a special group reflection tool that he can use to arrange an interactive after-action review process (Othlinghaus and Hoppe, 2016). The tool allows him to show and discuss important sequences from chat conversations of several players, review specific actions and aspects of their communication behavior. The data provided by the tool can be used to give feedback to the players and initiate group discussions to help them reflecting their actions and improving their performance.

Game Design

The game includes three different scenarios. The scenarios differ according to the type of customer used, especially in terms of conversation style (Rahim and Bonoma, 1979) and the problem situation of the customer, and thus in the level of difficulty. The first scenario serves as a base level and tutorial. The customer's problem can be solved quite easily by the player, since the conversation is reduced to the conversation phases in which only information content has to be collected and no pure "soft skill" phases have to be passed through. The conversation therefore only includes the greeting phase, the problem-solution phase and the conclusive phase. The customer in this scenario can be classified as an *integrating* customer according to the model of Rahim and Bonoma (Rahim and Bonoma, 1979), who differentiated five different styles of handling interpersonal conflicts, and is therefore open to reach a solution acceptable for both parties and is showing problem-solving behavior. According to a study conducted by Cho et al. (2002), the customer's problem is the third most common cause of non-public online customer complaints: delivery problems. The aim of this scenario is to help the player becoming acquainted with the user interface and let him walk through the basic milestones of the complaint conversation.

In the second scenario, the level of difficulty is increased. The customer is emotionally aroused because of his problem and must be calmed down. According to the classification of Rahim and Bonoma (1979), this customer is considered a *compromising* customer. The customer's problem is the most common problem within non-public online complaints (Cho et al., 2002): he has (among other things) problems with the customer service. The individualization on the delivered product is wrong and in a previous attempt to complain, the customer did not achieve a satisfactory result because there was a misunderstanding between the customer and the other member of the support staff. This makes the customer also a *follow-up complainant*, as it is the second time that he has contacted the customer service about the same problem (Stauss and Seidel, 2010). The result of the scenario is that after retrieving the database, the player learns that an error in production caused the incorrect individualization. Possible solutions to the problem in this scenario are replacement with the correct product or a refund. The goal of the scenario is to successfully pass through all five phases of a complaint process.

The third scenario is the one with the highest level of difficulty. Unlike the previous scenarios, it is less about processing the information milestones than about showing patience and applying soft skills. This customer can be classified as a *dominating* customer (Rahim and Bonoma, 1979) and a grouser (Stauss and Seidel, 2010). He is only focused on his own needs and shows little or no understanding for the other side. He tries to force a solution that is optimal for him and is looking for a continuation of the conflict. He has problems with the

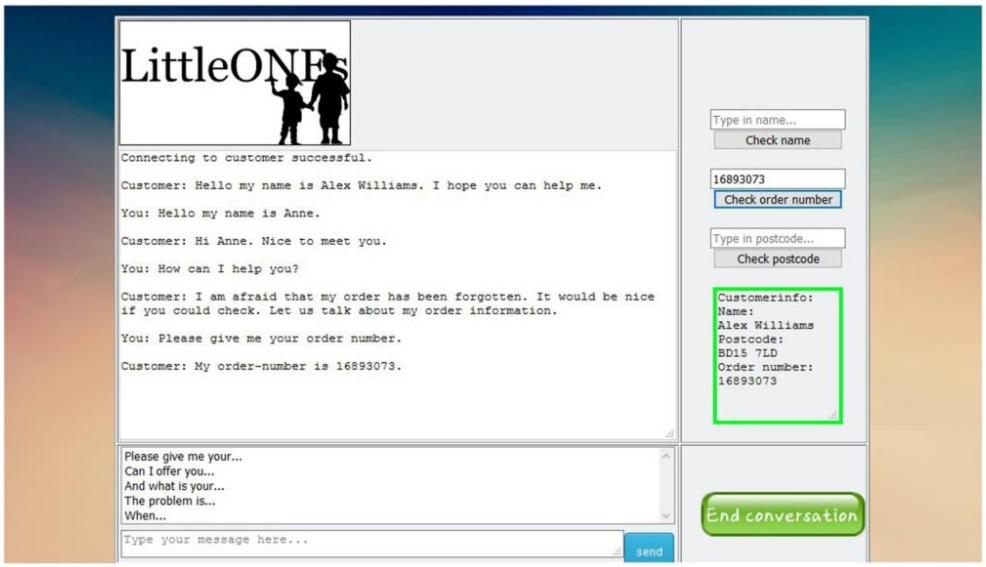

FIGURE 4 | Basic Interface of CuCoMaG.

business rules and conditions, which is the second most common problem with non-public online complaints according to Cho et al. (2002). The costumer is not reasonable and reacts abusive. When checking the database, the player learns that the customer is regularly complaining. The player's best result may be to not respond to the customer's provocations and finally end the conversation, since in this scenario, the player is not able to step into the next phase of the complaint process. This is called *active farewell* (Stauss and Seidel, 2010). The goal of this scenario is to deal with extreme situations. In this scenario, the player has to prove his ability to deal with provocations and difficult customers.

Evaluation Results

The playability, game experience, as well as the perception of the serious game itself and the interaction with the chatbot in particular have been evaluated in a mixed method study combining qualitative and quantitative methods (Othlinghaus-Wulhorst et al., 2019). The results of this study will be summarized and discussed in this section.

To investigate, whether the scenarios are perceived as realistic, if the developed scenarios' chatbots behave as intended, and whether their style of conversation is influencing how the players experience the chatbots, three hypotheses have been formulated (Othlinghaus-Wulhorst et al., 2019)—mainly relying on the game experience and perception of the chatbots (subjective measures) and evaluation of the chat transcripts (objective measures):

1. *"Participants who play the second scenario ("compromising") achieve different results in the game experience questionnaire (GEQ) dimensions tension, negative affect, and challenge than participants who play the third scenario ("dominating")."*
2. *"Participants who play the second scenario ("compromising") achieve different results in the Holtgraves questionnaire dimensions comfortable, thoughtful, polite, responsive, and engaging than participants who play the third scenario ("dominating")."*
3. *"Participants with prior experience/knowledge in complaint management achieve better results than participants without prior experience/knowledge."*

20 subjects (average 26.05, SD = 7.99, 15 females, 5 males) participated in the study. of the subjects indicated that they had prior experience in customer complaint management. All participants of the study played two scenarios, either the first ("integrating customer") and the second ("compromising customer"), or the first and the third ("dominating customer"). The distribution was randomized. Subsequent to the gaming session, the participants were asked to answer several post-experiment questionnaires to collect their experiences and perceptions during the game: (1) The *Game Experience Questionnaire* (GEQ) (Ijsselsteijn et al., 2013), (2) a questionnaire for the evaluation of educational role-playing games (Dell'Aquila et al., 2017), and (3) a questionnaire for measuring the human-like qualities of the chatbots developed by Holtgraves et al. (2007). In addition to the questionnaires and a subsequent

qualitative interview, the chat transcripts have been evaluated in regard to the answer quality of the chatbot. Based on human coding, every answer of the chatbot during the gaming session was assigned to one of three categories: *constructive*, *comprehensive*, and *nonsensical*. In order to estimate which of the predefined sentence openers have been used frequently, rarely or not at all, the frequency of uses for each one was counted.

The first hypothesis could only be partially confirmed. There were only significant differences in the dimension *negative affect*, but not in the dimensions *tension* and *challenge*. The lack of significant results could be possibly caused by methodical conditions. First, the number of participants was rather small. Second, the participants were asked to evaluate the perception of both played scenarios combined. The second hypothesis could be partially confirmed as well. Players who played the second scenario indeed showed significant differences in the dimensions *thoughtful*, *polite*, *responsive*, and *engaging*, but not in the dimension *comfortable*. As predicted, there were no significant differences in the dimensions *human* and *skilled*. This suggests that there is a difference in the style of conversation but not in the quality of the chatbots' implementation. In order to be able to examine hypothesis 3, we needed to define "success in the game." It has been determined by (a) a relative score calculated by the system and (b) the total number of inputs, as it was assumed that fast completion is an indicator for effective complaint management. Unfortunately, this hypothesis could not be tested due to the small sample size ($n = 4$).

The analysis of the chat transcripts revealed that *constructive* chatbot responses were the ones occurring most often, followed by *comprehensible* responses. The number of *nonsensical* responses has been quite low for all scenarios, which underlines the quality of the chatbot scripts. Reponses that were categorized as comprehensible, but not constructive, were default outputs, which were implemented for every sentence opener in case the free text part of the chat message was not understood by the chatbot. This way, the chatbot was able to show that he still understands the general gist or intention of the message. Sentence openers used to obtain information from the customers (e.g., "Tell me…", "Please describe…") were used most frequently, as well as the sentence opener "I am sorry…", which is not surprising, since apologies are almost always suitable in the given situation and clearly associated with polite behavior. In general, the results of the study showed that the idea and approach of the game were rated positively, but the evaluation also revealed problems, e.g., with the use of the sentence openers. It could be validated that the chatbots' style of discussion is influencing the players' perception of them, which emphasizes the successful design of the dialog scripts.

DIMENSIONS OF THE DESIGN OF SERIOUS ROLE-PLAYING GAMES FOR THE TRAINING OF SOCIAL SKILLS

The previous chapter has assembled examples instances of serious role-playing environments and ensuing empirical studies. In this chapter, we combine and inter-relate this experience with general issues in the design of serious role-playing games to devise and propose a set of design dimensions that constitute important aspects for the conceptualization, description, and comparison of serious role-playing games.

Learning Context

Many practical considerations have to be taken into account when designing serious role-playing games for the training of social skills. The probably most important is (as it is the case for every educational game) to have a clear educational purpose for using them (Whitton and Hollins, 2008). Digital games have a great motivational potential, but this potential needs to be utilized to convey the pedagogical goals and learning objectives. The goal of the game should be aligned to the learning outcomes as much as possible, otherwise learners may learn something, but it may not be what was intended. Learning objectives and intents need to be translated into concrete mechanical elements of gameplay by mapping learning mechanics and game mechanics onto each other. In games that pursue the goal to impart and train certain skills, learners should be given the opportunity to put these skills into practice in order to facilitate skills acquisition and provide a context in which these skills are useful (Naido et al., 2000).

Furthermore, the setting of the game needs to be appropriate for the learning context (Whitton and Hollins, 2008). As many studies mentioned in chapter Related Work show, role-playing games are an ideal instrument for the assessment and training of soft skills. However, the chosen scenario and storyline need to be appropriate in the given thematic context and should be described adequately for the players, so they are able to develop immediate understanding and empathy with the role they are assigned. The storyline may be fictitious, but the concepts used in it should be real to ensure that a transfer to real-word settings is possible (Pivec, 2009). Also, the desired learning outcome will not be achieved unless the correct game situation is chosen for the selected topic (Salen and Zimmerman, 2004). Another important point is that the educational design must be based on an underlying corpus of background theories. This includes general psychological and pedagogical concepts and guidelines for the design of serious games, as well as theoretical foundations of the learning material itself.

Technical Architecture and Set-Up

From the technical perspective, major issues regarding the implementation of serious role-playing games are flexibility, reusability, and extensibility/adaptability. Thanks to the use of a multi-agent blackboard architecture, our framework for scenario-based game development can easily be adapted and tailored to different settings and use cases, while the web-based gaming environment ensures easy access and platform independence. To adapt the framework to a new scenario, the following elements are needed: (1) a new GUI including sentence openers to be provided in menu selection, (2) new AIML scripts, and (3) modified or additional agents in the backend. The actual effort of course depends on the expertise of the developer.

The use of chatbots in serious role-playing games entails some major challenges. There are different approaches and technologies for natural language processing, each coming with specific advantages and disadvantages. In our approach, we use AIML as technological basis for the dialog modeling of the chatbots, but of course, there are many more approaches (e.g., data-driven technologies).

Dialog Models and Degrees of Freedom in Communication

There is a range of possibilities for introducing dialogs with virtual characters (not necessarily chatbots) in digital role-playing games. Brusk and Björk summarize different dialog models in games (Brusk and Björk, 2009): In some games, dialogs are the only way of interacting with the game, meaning that the dialog *is* the gameplay. In other games, dialogs are integrated as separate modes. Either they are taking place concurrently to other actions, or solely with no other activity occurring at the same time. In our scenarios, dialog is indeed the main gameplay element. There may be side tasks, but the focus is on the communication behavior of the players and their interaction with other characters (chatbots and/or other players, depending on the setting of the game).

In dialog-centric role play settings, one major design decision is related to the degrees of freedom in communication the players have. There is a range of communication models from fully predefined single choice inputs to free text composition. The choice is mainly depending on the setting, the narrative structure of the game, and the technical implementation. Using single choice inputs within underlying conversation trees are rather easy to implement, but provide the least freedom in communication. The players always have to select a predefined answer from a given set, and have no possibility to express themselves. Also, the game plot and the structure of the dialog is predefined. There may be decision nodes in the communication tree allowing for different lines of action, but the freedom of choice is very limited.

In our approach, we decided to integrate chatbots as dialog partners for the players. As illustrated in chapter Multi-Agent Architecture, it is very hard for natural language processing artificial intelligence to really grasp the sense of what has been said and a sophisticated chatbot design and implementation is a complex task. Thus, free text input poses a big challenge for developers. The use of sentence openers appears to be a compromise between these two ends of the spectrum. On the one hand, it limits the possible inputs, which reduces the complexity of the AIML scripts immensely and helps the chatbot to understand the general gist of a text input. On the other hand, it still offers the players the possibility to formulate their own inputs and express themselves more freely.

Feedback and Scaffolding Elements and Mechanisms

As we have shown in chapter Adaptive Feedback, feedback is crucial for ensuring the success of any serious role-playing game. It allows the learners to reflect on what happened during the role play and to analyze the consequences of their actions. In our approach, we differentiate between *ingame* and *aftergame* feedback. Ingame feedback refers to implicit feedback during the role-playing session. We realize this kind of feedback mainly through the reactions of the chatbots. Other feedback mechanisms are conceivable, but they should not corrupt or break the immersion during the role play situation. The balance between keeping the realism and immersion on the one hand and providing information on the status of the conversation as well as the players' performance and progress on the other hand is proposing a major design challenge for this kind of games. In our approach, aftergame feedback is an important point for enabling reflection processes. We consider a combination of an overall summary presented after the role play session and some kind of augmented replay of the dialog particularly helpful and promising.

Another important challenge for research and development in the area of serious role-playing games is to establish intelligent mechanisms for support and guidance (scaffolding). Learners should be provided with appropriate support in order to enable them to master the challenges of the game and achieve the learning goals. Ideally, a serious game should also adapt to the learners' level of knowledge, skills, as well as progress and current performance, as adaptation and personalization are considered key factors for education (Bellotti et al., 2010). Kickmeier-Rust and Albert suggest the introduction of micro-adaptive interventions (Kickmeier-Rust and Albert, 2010). This approach allows for interventions, support, guidance or feedback in a meaningful and personalized way, embedded in the game flow. These adaptive educational mechanisms are supposed to support the learner by hinting or providing appropriate feedback in certain situations, e.g., when misconceptions occur or when the progress is unsatisfactory (Kickmeier-Rust and Albert, 2010). The idea is to provide help to the learners by intelligently monitoring and interpreting their behavior in a non-invasive manner, which we consider a very promising approach. At this point of time, scaffolding, adaptation and personalization are incorporated in our framework only to a limited extend, thus augmenting these dimensions in our approach proposes a significant challenge for future research.

Relation Between Immersion and Reflection

As described in chapter Immersion and Reflection, one major advantage of games is their motivational and immersive potential. Immersion holds the potential to motivate learners and make them get more engaged in learning task and this potential needs to be used to full capacity in the role play situation. Getting immersed in a game requires some degree of (perceived) realism, because if learners do not perceive a scenario as realistic, they are likely to regard the game experience as irrelevant to their understanding of the real world (Sutcliffe, 2002). Thus, realism is an important characteristic of any successful serious role-playing game design. Ribbens and Malliet identified seven factors of perceived game realism: (1) simulation realism, (2)

freedom of choice, (3) character involvement, (4) perceptual pervasiveness, (5) authenticity regarding subject matter, (6) authenticity regarding characters, and (7) social realism (Ribbens and Malliet, 2010).

A properly designed serious role-playing game also needs to provide support for reflection, allowing the learners to re-think and reflect on their actions. There are approaches claiming that it could be beneficial to have reflection taking place within the game itself without letting the learner step out of the game world by offering reflection activities within the game (Yusoff et al., 2009). However, as we have argued in chapter Immersion and Reflection, there is reason to assume that immersion tends to hinder the critical self-reflection, and based on this assumption, we decided to separate the actual role play phase from the reflection phase in our framework, allowing the learners to step out of the game world and their role and take over a distant perspective during the reflection phase. Following Malzahn et al. (2010), we claim that reflection needs role distance, which is not compatible with a high degree of immersion (although this is desirable during the actual role play). Accordingly, phases of enactment (role play) should be separated from reflection. Reflection phases should enable to take a third-person perspective on the prior experience, which requires an accessible/readable representation of this experience. During this phase, immersion is explicitly undesirable in order to help learners to view their own actions from the perspective of an external observer.

Collaboration Support

An increasing popularity of multi-user virtual environments and games is causing a growing interest in the use of collaborative technologies for learning scenarios and recent research is indicating the positive effects of collaborative learning (Whitton and Hollins, 2008). Collaborative learning in the context of games describes a learning situation in which more than one learner participates in a learning (game) activity pursuing a common goal (Romero et al., 2012). In collaborative scenarios, learners work together on a common goal, they share and construct a certain level of knowledge, expertise and understanding (Romero et al., 2012). Major pedagogical benefits of bringing collaborative elements in gaming environments are (among others) providing multiple perspectives, creating self-awareness of the learning process, and thus making learning authentic and relevant (Whitton and Hollins, 2008). Serious games can provide a context for solving tasks and learning together with others. Integrating collaborative elements in a serious game may increase the players' motivation and foster the development of cognitive skills (Romero et al., 2012). In addition, collaborative virtual environments allow for a detailed recording of all collaborative interactions and thus may help to get a better understanding of those (Dillenbourg, 1999). Dillenbourg claims that it should be the aim of research to determine under which conditions collaborative learning is efficient (Dillenbourg et al., 1996).

He identifies three main criteria for rich and successful collaborative learning interactions (Dillenbourg, 1999): interactivity, synchronicity and negotiability. *Interactivity* is an integral part of any collaborative situation. It is not the frequency of interactions that defines the degree of interactivity, but the extent to which the interactions influence the other persons' cognitive processes. *Synchronicity* means that persons involved in a collaborative situation wait for messages from others and process them immediately. *Negotiability* relates to the structure of collaborative dialog being more complex than a hierarchical situation. That means one person will not impose her view only based on her authority, but will (to a certain extent) argue for her standpoint, justify, negotiate, and try to convince.

We have provided an example of a collaborative scenario in chapter Case Study: Conflict Management., in which two human players are involved in a mediation talk moderated by an AI-controlled mediator. However, not in all scenarios it is desirable and reasonable to include collaborative elements. It always depends on the context, the scenario and the learning objectives. If a task can be solved by one player, there is no need for collaboration. Thus, the tasks incorporated in the game should be only solvable if players act together and there should be a common goal (Wendel et al., 2013).

CONCLUSION

In this article, we presented a technical and conceptual framework for serious role-playing games for the training of specific social skills in virtual learning environments involving chatbots in dialog-centric settings. From the design perspective, three distinctive conceptual features characterize our framework: (1) chat-like interaction with an AI-controlled chatbot, (2) phases of immersion (role-playing) and reflection are separated to facilitate a change of perspective that is considered conducive for learning, and (3) the learning process is emphasized by means of adaptive feedback based on individual analyses. The technical conception is based on three main components: (1) AI-controlled chatbots that adapt to the player's behavior, (2) a multi-agent blackboard system as the backbone in order to keep components independent and to optimize performance due to parallel processing; and (3) intelligent support for an automated evaluation of the player's performance and feedback generation.

Different use cases based on this framework have been presented, including scenarios for the training of workplace-oriented conflict management, patient-centered medical interviews, and customer complaint management. First evaluation studies indicate that this approach is assessed positively, the scenarios are perceived as useful and realistic and may qualify for real training situations. Due to the flexible architecture, our framework can easily be tailored to different settings and use cases and thus serve as a basis for future research focusing on the adaptation to other contexts and systems.

Our framework facilitates the building of serious virtual role playing games in that it allows for tailoring and adapting a given component architecture with very limited effort, comprising the provision of a specific GUI with sentence openers, a new set of AIML scripts (chatbots), and (possibly) a modification/extension of the backend agents. The framework provides all the basic mechanisms such as the inter-operability between GUI, chatbots, and agents through a tuple space. The basic architecture is

available as a kind of modifiable prototype. Other than model-driven development (Schmidt, 2006), our approach does not use meta-level descriptions in combination with generators in the overall systems engineering process. Only the AIML-based specification of chatbot behavior could be conceived as a meta-level element. However, this is limited to one of the components and only imported and exploited in our application framework. Although our system architecture and basic mechanisms are predefined, these premises do not preclude the quality of the ensuing application instances. These depend very much on the specification of chatbot scripts as well as on the GUI design. Accordingly, our evaluations have relied on standard instruments to measure game experience and usability as the main human-oriented factors.

Based on our experience, we formulated a set of general dimensions and challenges in the design of serious role-playing games for the training of social skills. In summary, we identified six major aspects: The *learning context* builds the basis of each serious game and relates to its theoretical foundation and the desired learning outcomes. The *technical architecture and set-up* refer to technologies and tools that are used for the technical implementation of such games and the underlying system architecture. *Dialog models and degrees of freedom in communication* relates to the question of how the communication with the non-playing dialog partner(s) is carried out, structured, and controlled, from predefined answers to sentence openers to free text input. *Feedback and scaffolding elements and mechanisms* are essential for the transfer of learning to application in the real world and can be integrated in many different ways. The *relation between immersion and reflection* refers to the question whether phases of immersion and reflection overlap or occur separate from each other. *Collaboration support* relates to the number of (human) players involved in the game and the question whether it enables collaborative learning.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

JO-W has been involved in a guiding role (main responsibility) in the design, implementation, and evaluation of the serious role-playing games ColCoMa and CuCoMaG described in chapter Case Studies. JO-W and HH are (co-)authors of the relevant publications connected to these developments. JO-W wrote the main manuscript with text input from HH. All authors have reviewed and approved the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank several students and employees of the University of Duisburg-Essen who were involved in the design, development, evaluation, and publication of the mentioned use cases, namely Katharina Emmerich, Katja Neuwald, and Anna Jedich (ColCoMa), Philipp Behler, Ingo Börsting, Heike Choi, Evelyn Fricke, Stefan Liszio, Christian Klöpfel, and Anna Kizina (training scenario for medical interviews), as well as Dorian Doberstein, Nadja Agreiter, Marco Bäumer, Menglu Cui, Shaghayegh Abdollahzadegan, Diba Heidari, Nan Jiang, Markus Mentzel, Huangpan Zhang, Hao Zheng, Viet Hung Dinh, and Anne Mainz (CuCoMaG).

REFERENCES

Abdul-Kader, S. A., and Woods, J. (2015). Survey on chatbot design techniques in speech conversation systems. *Int. J. Adv. Comput. Sci. Appl.* 6, 72–80. doi: 10.14569/IJACSA.2015.060712

Annetta, L. A. (2010). The "I's" have it: a framework for serious educational game design. *Rev. Gen. Psychol.* 14, 105–113. doi: 10.1037/a0018985

Augello, A., Gentile, M., and Dignum, F. (2016). "Social agents for learning in virtual environments," in *International Conference of Games and Learning Alliance* (Cham: Springer), 133–143.

Aylett, R. S., Louchart, S., Dias, J., Paiva, A., and Vala, M. (2005). "FearNOT! - an experiment in emergent narrative," in *International Workshop on Intelligent Virtual Agents* (Berlin; Heidelberg: Springer), 305–316.

Beaumont, R., and Sofronoff, K. (2008). A multi-component social skills intervention for children with asperger syndrome: the junior detective training program. *J. Child Psychol. Psychatry* 49, 743–753. doi: 10.1111/j.1469-7610.2008.01920.x

Behler, P., Börsting, I., Choi, H., Fricke, E., Liszio, S., Klöpfel, C., et al. (2013). "Eigentlich geht es mir gut - Entwicklung eines Serious Games zur patientenzentrierten Gesprächsführung," in *DeLFI 2013: Die 11. E-Learning Fachtagung Informatik*, eds A. Breiter and C. Rensing (Bonn: Gesellschaft für Informatik e.V.), 11–23.

Bellotti, F., Berta, R., De Gloria, A., and Primavera, L. (2010). Supporting authors in the development of task-based learning in serious virtual worlds. *Br. J. Educ. Technol.* 41, 86–107. doi: 10.1111/j.1467-8535.2009.01039.x

Bezuijen, A. (2012). *Teamplay: the further development of TeamUP, a teamwork focused serious game* [Master's thesis]. Delf University of Technology, Delft, Netherlands.

Bosse, T., and Gerritsen, C. (2016). "Towards serious gaming for communication training - a pilot study with police academy students," in *International Conference on Intelligent Technologies for Interactive Entertainment* (Cham: Springer), 13–22.

Boud, D., Keogh, R., and Walker, D. (1985). "Promoting reflection in learning: a model," in *Reflection: Turning Experience Into Learning*, eds D. Boud, R. Keogh, and D. Walker (London: Kogan Page), 18–40.

Brusk, J., and Björk, S. (2009). "Gameplay design patterns for game dialogues," in *Proceedings of the 2009 Digital Games Research Association Conference* (London).

Cho, Y., Im, I., Hiltz, R., and Fjermestad, J. (2002). "An analysis of online customer complaints: implications for web complaint management," in *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* (Big Island, HI: IEEE), 2308–2317. doi: 10.1109/HICSS.2002.994162

Combs, M. L., and Slaby, D. A. (1977). Social-skills training with children. *Adv. Clin. Child Psychol.* 1, 161–201. doi: 10.1007/978-1-4613-9799-1_5

Corti, K. (2006). Games-based Learning; a serious business application. *Informe PixelLearn*. 34, 1–20. Available online at: http://www.pixelearning.com/docs/seriousgamesbusinessapplications.pdf

De Freitas, S., and Oliver, M. (2006). How can exploratory learning with games and simulations within the curriculum be most effectively evaluated? *Comput. Educ.* 46, 249–264. doi: 10.1016/j.compedu.2005.11.007

Dell'Aquila, E., Marocco, D., Ponticorvo, M., Di Ferdinando, A., Schembri, M., and Miglino, O. (2017). *Educational Games for Soft-Skills Training in Digital Environments*. Cham: Springer. doi: 10.1007/978-3-319-06311-9

Di Ferdinando, A., Schembri, M., Linchan, C., Linehan, C., and Miglino, O. (2011). *Learn to Lead - A Web Based Game to Teach Leadership Theories in Vocational Courses. Games and Creativity in Education and Training*. Naples: Fridericiana Editrice Universitaria.

Dillenbourg, P. (1999). What Do You Mean by Collaborative Learning? *Collaborative Learning: Cognitive and Computational Approaches*. Oxford: Elsevier, 1–19.

Dillenbourg, P., Baker, M., Blaye, A., and O'Malley, C. (1996). The Evolution of Research on Collaborative Learning. Learning in Humans and Machine: Towards an Interdisciplinary Learning Science. Oxford: Elsevier, 189–211.

Doberstein, D., Agreiter, N., Bäumer, M., Cui, M., Abdollahzadegan, S., Heidari, D., et al. (2016). "CuCoMaG - group reflection support in role-playing environments," *in DeLFI 2016 - Die 14. E-Learning Fachtagung Informatik*, ed U. Lucke et al. (Bonn: Gesellschaft für Informatik e.V.), 327–329.

Emmerich, K., Neuwald, K., Othlinghaus, J., Ziebarth, S., and Hoppe, H. U. (2012). "Training conflict management in a virtual environment," *International Conference on Collaboration and Technology* (Berlin; Heidelberg: Springer), 17–32.

Even, C., Bosser, A. G., Ferreira, J., Buche, C., Stéphan, F., Cavazza, M., et al. (2016). "Supporting social skill rehabilitation with virtual storytelling," in *Proceedings of the 29th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2016)* (Key Largo, FL), 329–334.

Festinger, L. (1954). A theory of social comparison processes. *Hum. Relat.* 7, 117–140. doi: 10.1177/001872675400700202

Flynn, R., McKinnon, L., Bacon, E., and Webb, J. (2011). "Maritime city: using games technology to train social workers - some initial results," in *International Conference on Entertainment Computing*. (Berlin; Heidelberg: Springer), 415–418.

Gebhard, P., Schneeberger, T., André, E., Baur, T., Damian, I., Mehlmann, G., et al. (2018). Serious games for training social skills in job interviews. *IEEE Trans. Games* 11, 340–351.

Gee, J. P. (2007). *Good Video Games and Good Learning: Collected Essays on Video Games, Learning, and Literacy*. New York, NY: Lang.

Gelernter, D. (1985). Generative communication in Linda. *ACM Transac. Prog. Lang. Syst.* 7, 80–112. doi: 10.1145/2363.2433

Glock, F., Junker, A., Kraus, M., Lehrian, C., Schäfer, A., Hoffmann, S., et al. (2011). ""Office Brawl" - A Coversational Storytelling Game and its Creation Process," in *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology (ACE 2011)* (New York: ACM).

Gresham, F., and Elliott, S. N. (2008). *Social Skills Improvement System (SSIS) Rating Scales*. Bloomington, IN: Pearson Assessment.

Gunter, G. A., Kenny, R. F., and Vick, E. H. (2006). A case for a formal design paradigm for serious games. *J Int. Dig. Media Arts Assoc*, 3, 93–105. Available online at: https://www.academia.edu/download/49480493/Gunter_20Kenny_20Vick_20paper.pdf

Haferkamp, N., Kraemer, N., Linehan, C., and Schembri, M. (2011). Training desaster communication by means of serious games in virtual environments. *Entertain. Comput.* 2, 81–88. doi: 10.1016/j.entcom.2010.12.009

Harteveld, C. (2011). *Triadic Game Design*. London: Springer.

Holtgraves, T. M., Ross, S. J., Weywadt, C. R., and Lin, H. T. (2007). Perceiving artificial social agents. *Comput. Hum. Behav.* 23, 2163–2174. doi: 10.1016/j.chb.2006.02.017

Hubal, R. C., Frank, G. A., and Guinn, C. I. (2003). "Lessons learned in modeling schizophrenic and depressed responsive virtual humans for training," in *Proceedings of IUI* (Miami, FL), 85–92. doi: 10.1145/604045.604062

Hunicke, R., Leblanc, M., and Zubek, R. (2004). "MDA: a formal approach to game design and game research," in *Proceedings of the Challenges in Games AI Workshop, 19th National Conference of Artificial Intelligence* (San Jose, CA: AAAI Press), 1-5.

Ijsselsteijn, W. A., De Kort, Y. A., and Poels, K. (2013). *The Game Experience Questionnaire*. Eindhoven: Technische Universiteit Eindhoven.

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., et al. (2008). Measuring and defining the experience of immersion in games. *Int. J. Hum. Comput. Stud.* 66, 641–661. doi: 10.1016/j.ijhcs.2008.04.004

Jeuring, J., Grosfeld, F., Heeren, B., Hulsbergen, M., Ijntema, R., Jonker, V., et al. (2015). *Communicate! - A Serious Game for Communication Skills. EC-TEL 2015*. Cham: Springer, 513–517.

Johnson, C. I., Bailey, S., and Van Buskirk, W. L. (2017). "Designing effective feedback messages in serious games and simulations: a research review," in *Instructional Techniques to Facilitate Learning and Motivation of Serious Games*, eds P. Wouters and H. van Oostendor (Cham: Springer), 119–140.

Jonassen, D., Mayes, T., and McAleese, R. (1993). "A manifesto for a constructivist approach to uses of technology in higher education," in *Designing Environments for Constructive Learning*, eds T. M. Duffy, J. Lowyck, and D. H. Jonassen (Berlin; Heidelberg: Springer), 231–247.

Kenny, R., and Gunter, G. (2011). Factors affecting adoption of video games in the classroom. *J. Interact. Learn. Res.* 22, 259–276.

Kerly, A., Hall, P., and Bull, S. (2006). "Bringing chatbots into education: Towards natural language negotiation of open learner models," in *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (London: Springer), 179–192.

Kesti,M. O., Leinonen, J., and Kesti, T. (2017). "The productive leadership game: fromtheory to games-based learning," in *Public Sector Entrepreneurship and the Integration of Innovative Business Models*, eds M. Lewandowski and B. Kozuch (Hershey: IGI Global), 238-260. doi: 10.4018/978-1-5225-2215-7.ch010

Kickmeier-Rust, M. D., and Albert, D. (2010). Micro-adaptivity: protecting immersion in didactically adaptive digital educational games. *J. Comput. Assist. Learn.* 26, 95–105. doi: 10.1111/j.1365-2729.2009.00332.x

Knode, S., and Knode, J.-D. (2011). "Using a simulation program to teach leadership," in *Proceedings of the 2011 ASCUE Summer Conference* (North Myrtle Beach, SC), 86–92.

Kortmann, R., and Harteveld, C. (2009). "Agile game development: lessons learned from software engineering," in *Proceedings of the 40th Annual Conference of the International Simulation and Gaming Association* (Singapore).

Kruse, J. (2000). *Diagnostische Falleinschätzung bei Patienten mit Psychischen und psychosomatischen Beschwerden und Störungen in hausärztlichen Praxen* [Professorial dissertation]. Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany.

Lane, H. C., and Hays, M. J. (2008). "Getting down to business: teaching cross-cultural social interaction skills in a serious game," in *Workshop on Culturally Aware Tutoring Systems (CATS)* (Montreal, QC), 35–46.

Leary, T. (1957). *Interpersonal Diagnosis of Personality: Functional Theory and Methodology for Personality Evaluation*. Oxford: Ronald Press.

Lim, M. Y., Aylett, R., Enz, S., Kriegel, M., Vannini, N., Hall, L., et al. (2009). Towards intelligent computer assisted educational role-play. *Edutainment* 2009, 208–219. doi: 10.1007/978-3-642-03364-3_27

Linssen, J., de Groot, T., Theune, M., and Bruijnes, M. (2014). "Beyond simulations: serious games for training interpersonal skills in law enforcement," in *Proceedings of European Social Simulation Association*, 604–607. Available online at: http://ddd.uab.cat/record/125597

Malzahn, N., Buhmes, H., Ziebarth, S., and Hoppe, H. U. (2010). "Supporting reflection in an immersive 3D learning environment based on role-play," in *European Conference on Technology Enhanced Learning* (Berlin; Heidelberg: Springer), 542–547.

Marocco, D., Pacella, D., Dell'Aquila, E., and Di Ferdinando, A. (2015). "Grounding serious game design on scientific findings: the case of ENACT on soft skills training and assessment," in *Design for Teaching and Learning in A Networked World (9307)*, eds G. Conole, C. Klobučar, J. Rensing, J. Konert, and É. Lavoué (Switzerland: Springer), 441–446.

Marr, A. C. (2010). *Serious Games für die Informations- und Wissensvermittlung. Bibliotheken auf Neuen Wegen*. Wiesbaden: Dinges & Frick.

Martens, A., Diener, H., Malo, S., Pan, Z., Cheok, A., Müller, W., et al. (2008). Games-based learning with computers - learning, simulations, and games. *Transac. Edutain* 1, 172–190. doi: 10.1007/978-3-540-69744-2_15

Mateas, M., and Stern, A. (2003). "Façade: an experiment in building a fully-realized interactive drama," in *Proceedings of the Game Developers Conference (GDC '03)* (San Jose,CA).

Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., van Ruijven, T., et al. (2014). The research and evaluation of serious games: Toward a comprehensive methodology. *Br. J. Educ. Technol.* 45, 502–527. doi: 10.1111/bjet.12067

McMahon, M., and Ojeda, C. (2008). "A model of immersion to guide the design of serious games," in *Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare and Higher Education* (Las Vegas), 1833–1842.

Memarzia, M., and Star, K. (2011). "Choices and voices - a serious game for prevent violent extremism," in *Intelligence Management: Knowledge Driven Frameworks for Combating Terrorism and Organized Crime*, eds S. Yates and B. Akhgar (Heidelberg: Springer), 130–142.

Michael, D., and Chen, S. (2006). *Serious Games: Games that Educate, Train, and Inform*. Boston, MA: Thomson Course Technology.

Mitchell, A., and Savill-Smith, C. (2005). *The Use of Computer and Video Games for Learning. A Review of the Literature*. London: LSDA.

Murray, J. H. (2017). *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. Cambridge; London: MIT press.

Naido, S., Ip, A., and Linser, R. (2000). Dynamic goal-based role-play simulation on the web: a case study. *J. Educ. Technol. Soc.* 3, 190–202. Available online at: https://www.jstor.org/stable/10.2307/jeductechsoci.3.3.190

Othlinghaus, J., and Hoppe, H. U. (2016). "Supporting group reflection in a virtual role-playing environment," in *International Conference on Intelligent Technologies for Interactive Entertainment* (Cham: Springer), 292–298.

Othlinghaus-Wulhorst, J., Jedich, A., Hoppe, H. U., and Harrer, A. (2018). "Using eye-tracking to analyze collaboration in a virtual role play environment," in *International Conference of Collaboration and Technology* (Cham: Springer), 185–197.

Othlinghaus-Wulhorst, J., Mainz, A., and Hoppe, H. U. (2019). "Training customer complaint management in a virtual role-playing game: a user study," in *European Conference on Technology Enhanced Learning* (Cham: Springer), 436–449.

Pivec, P. (2009). *Game-Based Learning or Game-Based Teaching*? British Educational Communications and Technology Agency (BECTA), Report No. 1509.

Prensky, M. (2000). *Digital Game-Based Learning*. New York, NY: McGraw-Hill.

Proksch, S. (2010). *Konfliktmanagement in Unternehmen. Mediation als Instrument für Konflikt- und Kooperationsmanagement am Arbeitsplatz*. Heidelberg: Springer.

Rahim, A., and Bonoma, T. V. (1979). Managing organizational conflict: A model for diagnosis and intervention. *Psychol. Rep.* 44, 1323–1244. doi: 10.2466/pr0.1979.44.3c.1323

Ribbens, W., and Malliet, S. (2010). Perceived digital game realism: a quantitative exploration of its structure. *Pres. Teleoperat. Virtual Environ.* 19, 585–600. doi: 10.1162/pres_a_00024

Ritterfeld, U., Cody, M., and Vorderer, P. (2009). *Serious Games: Mechanics and Effects*. New York, NY: Routledge.

Romero, M., Usart, M., Ott, M., Earp, J., and de Freitas, S. (2012). "Learning through playing for or against each other? Promoting collaborative learning in digital game based learning," in *Proceedings of the 20th European Conference on Information Systems (ECIS) 2012* (Barcelona). Available online at: http://aisel.aisnet.org/ecis2012/93

Salen, L., and Zimmerman, E. (2004). *Rules of Play: Game Design Fundamentals*. Cambridge; London: The MIT Press.

Schatzki, T. R. (1996). *Social Practices: A Wittgensteinian Approach to Human Activity and the Social*. Cambridge: Cambridge University Press.

Schmidt, D. C. (2006). Model-driven engineering. *IEEE Comput.* 39:25. doi: 10.1109/MC.2006.58

Schuster, R. J. (2010). Gruppenreflexion als Kommunikationsinstrument. Management und Wirtschaft. *Schriftenreihe zur Wirtschaftlichen Forschung*, 13, 7–23. Available online at: http://www.fhvie.ac.at/Forschung/Publikationen/Schriftenreihe

Schwantes, U., and Kampmann, M. (2007). Gesundheitsorientierte Gesprächsführung. *Der Mensch* 38. Available online at: https://www.dachverband-salutogenese.de/cms/der-mensch/heft-38-12007/

Shawar, B. A., and Atwell, E. (2007). Chatbots: are they really useful? *Ldv Forum* 22, 29–49.

Shute, V. J. (2008). Focus on formative feedback. *Rev. Educ. Res.* 78, 153–189. doi: 10.3102/0034654307313795

Squire, K., and Jenkins, H. (2003). Harnessing the power of games in education. *Insight* 3, 5–33.

Stauss, B., and Seidel, W. (2010). "Complaint management," in *Introduction to Service Engineering*, eds G. Salvendy and W. Karwowski (Hoboken, NJ: John Wiley & Sons), 414–432.

Susi, T., Johannesson, M., and Backlund, P. (2007). *Serious Games: An Overview. Technical Report HS-IKI-TR-07-001*. Skövde: School of Humanities and Informatics, University of Skövde.

Sutcliffe, M. (2002). "Simulations, games and role-play," in *Handbook for Economics Lecturers*, ed P. Davies (Bristol: The Higher Education Academy, The Economics Network), 1–26.

Totty, M. (2005). Better training through gaming. *Wall Street J. Eastern Ed.* 245:R6.

Vaassen, F., and Wauters, J. (2012). "deLearyous: training interpersonal communication skills using unconstrained text input," in *Proceedings of ECGBL* (Cork), 505–513.

Van Dijk, D., Dick, R., Hunneman, R., and Wildlevuur, S. (2008). Self City: "Training social skills in a game," in *Proceedings of Second European Conferences on Game-based Learning* (Academic Publishing Limited) (Barcelona), 481–488.

Walk, W., Görlich, D., and Barrett, M. (2017). "Design, dynamics, experience (DDE): an advancement of the MDA framework for game design," in *Game Dynamics* eds O. Korn and N. Lee (Cham: Springer), 27–45.

Wallace, R. (2004). *The Elements if AIML Style*. San Francisco, CA: ALICE AI Foundation.

Weinbrenner, S. (2012). *SQLSpaces - A Platform for Flexible Language-Heterogeneous Multi-Agent Systems* [Doctoral dissertation]. Verlag Dr. Hut, Munich, Germany.

Wendel, V., Gutjahr, M., Göbel, S., and Steinmetz, R. (2013). Designing collaborative multiplayer serious games. *Educ. Inform. Technol.* 18, 287–308. doi: 10.1007/s10639-012-9244-6

Whitton, N., and Hollins, P. (2008). Collaborative virtual gaming worlds in higher education. *ALT-J* 16, 221–229. doi: 10.3402/rlt.v16i3.10900

Winkler, R., and Söllner, W. (2018). *Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis*. Chicago, IL: Academy of Management Annual Meeting (AOM).

Yee, N. (2006). The demographics, motivations, and derived experiences of users of massively multi-user online graphical environments. *Pres. Teleoperat. Virtual Environ.* 15, 309–329. doi: 10.1162/pres.15.3.309

Yusoff, A., Crowder, R., Gilbert, L., and Wills, G. (2009). "A conceptual framework for serious games," in *2009 Ninth IEEE International Conference on Advanced Learning Technologies* (Riga: IEEE), 21–23. doi: 10.1109/ICALT.2009.19

Ziebarth, S., Kizina, A., Hoppe, H. U., and Dini, L. (2014). "A serious game for training patient-centered medical interviews," in *14th International Conference on Advanced Learning Technologies* (Athens: IEEE), 213–217. doi: 10.1109/ICALT.2014.69

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Othlinghaus-Wulhorst and Hoppe. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

3 CASE STUDY: CONFLICT MANAGEMENT

USING EYE-TRACKING TO ANALYZE COLLABORATION IN A VIRTUAL ROLE PLAY ENVIRONMENT

This paper was presented at the 24th International Conference on Collaboration and Technology (CRIWG) 2018 and published as part of the Lecture Notes in Computer Science book series (LNCS, volume 11001). The conference is a major forum for academic researchers to exchange their experiences related to the development and use of collaboration technology and has a strong focus on technology design and development.

Reprinted by permission from Springer Nature: Springer Nature, Collaboration and Technology. CRIWG 2018. Othlinghaus-Wulhorst, J., Jedich, A., Hoppe, H. U. & Harrer, A.: Using Eye-Tracking to Analyze Collaboration in a Virtual Role Play Environment. © Springer Nature Switzerland AG (2018)

Author	**Contribution**	**%**
Julia Othlinghaus-Wulhorst	- Main responsibility in the design, implementation and evaluation of the game - Conceptualization of the evaluation study - Writing of main manuscript	60%
Anna Jedich	- Conduction of the evaluation study - Statistical evaluation	15%
Andreas Harrer	- Research planning and guidance - Setting up of lab environment (for eye-tracking) - Text revision	15%
H. Ulrich Hoppe	- Research planning and guidance - Text revision	10%

Using Eye-Tracking to Analyze Collaboration in a Virtual Role Play Environment

Julia Othlinghaus-Wulhorst[1(✉)], Anna Jedich[1], H. Ulrich Hoppe[1], and Andreas Harrer[2]

[1] University of Duisburg-Essen, Lotharstraße 63, 47057 Duisburg, Germany
{othlinghaus,hoppe}@collide.info,
anna.jedich@stud.uni-due.de
[2] Dortmund University of Applied Sciences and Arts,
Emil-Figge-Str. 42, 44227 Dortmund, Germany
andreas.harrer@fh-dortmund.de

Abstract. The ColCoMa environment supports the training of workplace-oriented conflict management strategies through virtual role play. The role play relies on a web-based environment in which the participants interact through chat dialogues. Two of the participants (the parties in conflict) are human actors whereas the third role ("mediator") is occupied by a chatbot. Our study aims at exploring the potential of eye-tracking analyses to assess the quality of cooperation in this situation. The standard assumption is that a certain "convergence" of the visual foci of attention between cooperation partners indicates better coordination and consideration of the other party. In our scenario, this assumption has to be refined by taking into account the different roles (including the role of the chatbot) and the distribution of utterances on the chat history. The eye-tracking parameters are compared to quality criteria such as successful completion of the game or richness/mutuality of the chat interactions. There are quite strong correlations on the aggregate level (taking overall eye-tracking convergence as a global parameter), yet not in terms of synchronicity between convergent eye-tracking and chat interaction. This is possibly due to the specific distribution of roles in our virtual training environment.

Keywords: Eye-tracking · Collaboration · Role play

1 Introduction

Collaborative virtual environments are increasingly recognized as effective and powerful tools for supporting learning and training, and a multitude of contexts and pedagogical approaches have been supported by different types of collaborative virtual environments [1]. Virtual role play environments are a special form of serious games that provide mobile and repeatable settings for learners and allow them to take over roles in particular contexts and to learn from the enacted experiences [2, 3]. In a safe environment, they can act, experiment, learn and teach without risking irreversible consequences [2].

© Springer Nature Switzerland AG 2018
A. Rodrigues et al. (Eds.): CRIWG 2018, LNCS 11001, pp. 185–197, 2018.
https://doi.org/10.1007/978-3-319-99504-5_15

It is of specific interest to combine virtual role playing with the collaborative learning perspective [1]. Dillenbourg [4] identifies the following criteria as important ingredients of rich and successful computer-supported collaborative learning interactions: interactivity, synchronicity and negotiability. *Interactivity* needs to be part of any collaborative situation, while its degree is not defined by the frequency of interactions, but rather by the extent to which the partners' cognitive processes are influenced by those interactions. *Synchronicity* means that the persons involved in such a situation are waiting for messages from the others' and process them immediately after delivery, so it is less a technical parameter than a social rule. *Negotiability* refers to the structure of collaborative dialog as being more complex than hierarchical situations: a partner will (to some extent) argue for his standpoint, justify and try to convince instead of imposing his view on the sole basis of his or her authority. According to Dillenbourg, collaboration only works under certain conditions and it should be the aim of research to determine the conditions under which collaborative learning is efficient [5]. Collaborative virtual environments may help to gain a better understanding of the underlying mechanisms of collaborative interactions by enabling a detailed recording of all interactions on the one hand and a careful design of the empirical situation on the other hand [4].

Many studies of collaborative learning rely on the quantitative or qualitative analysis of dialogues (or multi-party conversations), i.e. they focus on verbal interactions. Eye-tracking, which has become more and more easily available and applicable in recent years, allows for enriching the analysis techniques towards non-verbal behavior. The possibility to track people's eye gaze can provide rich and insightful information and offer unique opportunities to understand their cognitive and perceptual processes [6]. Especially using several eye-tracking devices in parallel, combined with other types of analysis, is helpful to afford an indication of the level of gaze synchronicity of the different group members, allowing to measure to what extend one person is "with" the other, i.e. how much are persons looking at the same thing at the same time [6]. This approach of using eye-tracking and automated analysis as a complement to content analysis methods is a recent development in the field of learning analytics.

Our study provides interesting opportunities to understand collaboration using eye-tracking data by analyzing visual coordination and its relationship to collaborative learning by investigating the fundamental question if there is a correspondence between gaze synchronicity and the quality of collaboration. The usage of eye-tracking analysis within collaborative serious games is to our knowledge a novel contribution as well as the exploration of interdependencies between synchronicity of gazes and collaboration quality, which is related to the dimension interactivity mentioned above.

2 Related Work

In relation to studies on collaboration and collaborative learning, eye-gaze can serve as an analysis technique, yet mutual gaze awareness can also be used as an additional channel of interaction and not only as an analytic instrument. Both aspects have been the subject of recent studies:

Richardson and Dale investigated the coupling between a speaker's and a listener's eye movements [7]. In their observational study, one set of participants (the speakers) talked spontaneously about a television show whose characters were displayed on a screen in front of them. Later, the other set of participants (the listeners) listened to the recorded monologues while looking at the same visual scene. By using a cross-recurrence analysis, Richardson and Dale found out that a listener's eye movements most closely matched a speaker's eye movements at a delay of 2 s and the more closely a listener's eye movements were coupled with the ones of a speaker, the better the listener performed on a comprehension test. Thus, the results indicate that the coupling between the eye movements of speaker and listener reflects the success of their communication.

Cherubini et al. developed an algorithm for detecting misunderstanding in a remote collaboration, which combines a linguistic model with eye-tracking data [8]. The participants of their study had to collaborate remotely via a chat tool and a map. The algorithm was designed to detect misunderstandings between the two persons using their eye movements on the shared workspace, their utterances containing references on the plan, as well as the availability of explicit referencing. This proposal was based on the finding that participants look at the points they are talking about in their messages at above the level of chance, and that the eye movements are denser around these points compared to any other region looked at during the same time period during reading or editing messages containing references to the shared workspace. The algorithm associates the distance between the gaze of the emitter and the receiver of a message with the probability that the message was not understood by the recipient. The results of the study show that the likelihood of misunderstandings is increased, if there is more dispersion.

Jermann et al. used synchronized eye-trackers to assess how dyads of programmers worked collaboratively on a code segment [9]. They contrasted a 'good' and a 'bad' dyad, and the results of their work suggest that a productive collaboration is associated with high visual recurrence. Schlösser et al. have used both mutual gaze awareness as well as eye-tracking analysis in a study of collaborative problem solving. They found that gaze awareness had a positive effect on problem solving results. They also identified share gaze events as an indicator of collaboration quality [10].

Schneider and Pea performed an eye-tracking study on collaborative problem-solving dyads [6]. These dyads collaborated remotely via an audio channel to learn from contrasting cases involving basic concepts about how the human brain is processing visual information. While in one group, the dyads were able to see the eye gaze of their partner on the screen, in the control group, they had no access to the information on their partner's gaze. The results of the study indicate that this real-time mutual gaze perception intervention enhances collaborative learning and collaboration quality. Collaboration quality was rated using dimensions developed by Meier et al. [11], who developed a five-point scale across nine dimensions (sustaining mutual understanding, dialogue management, information pooling, reaching consensus, task division, task management, technical coordination, reciprocal interaction, and individual task orientation) for assessing collaboration.

Schneider and Pea combined joint visual attention, network analysis and machine learning to predict dimensions of productive collaboration [12] based on the dataset of

their previous study [9]. They visualized the eye-tracking data as networks, where the nodes of the graph represent fixations and edges represent saccades and used network metrics to interpret the properties of the graph to find proxies for the collaboration quality of the participants. They found that different characteristics of the graphs correlated with different dimensions of collaboration quality. Those characteristics have been used to predict the collaboration quality by using a machine-learning algorithm. This way it was possible to roughly predict the participants' collaboration quality with an accuracy between 85% and 100%. Here again the rating scheme for assessing collaboration quality by Meier et al. [11] has been used.

Sharma et al. applied the so-called *extreme value theory* (EVT) to eye-tracking data collected during an online collaborative problem-solving task in order to predict the collaboration quality [13]. In the univariate mode, each pair of time episodes from participants A and B is substituted by a measure of their differences, which results in a series of single values. In the bivariate mode, they also took into account the dynamic coupling of the two-time series. A comparison between the results by EVT and traditional approaches revealed that EVT provides a better prediction of the collaboration quality.

The results of the mentioned studies support the idea that eye gaze synchronicity is crucial for effective collaboration. In our study, we attempt to explore the dependencies between gaze synchronicity that is automatically computed and collaboration quality that is measured via a rating scheme manually, thus combining these two lines of analytical methods within one study.

3 Virtual Role Play Environment: ColCoMa

For our eye-tracking based study, we used the collaborative game-like environment *ColCoMa* (Collaborative Conflict Management), which supports the training of workplace-oriented conflict management strategies through virtual role play [13]. It engages two players to participate in a chat conversation about a given fictitious conflict in a 2D virtual environment. The conversation is moderated by a chat bot acting as mediator, and follows the typical structure of mediation talks. The main goal is to come to a conflict resolution at the end of the conversation by showing appropriate and constructive behavior.

In ColCoMa, each player is assigned a predefined role in the conflict scenario: Mr. Meier is working as a member of the computer support hotline team in a software company and conscientiously takes much time for his customers. His supervisor, Mrs. Schmidt, does not embrace Mr. Meier's very long call sessions. She wants him to work more efficiently. After a negative appraisal of Mr. Meier's performance on the part of Mrs. Schmidt, the situation escalates. The scenario is intentionally kept simple and comprehensible, focusing on the main conflict and each person's feelings in order to support both immediate understanding and empathy with the assigned role.

During the mediation, each player is seeing cartoon-like representations of the other conflict party and the mediator on screen, similarly to sitting opposite to the dialog partners. The players are able to communicate with each other and the mediator via an integrated chat. The players also have the possibility to evoke facial animations through

the character images by using common emoticons. Figure 1 is showing the general interface of the ColCoMa environment. Besides the chat section, it includes a notepad and a help section in order to get additional information about the scenario as well as the game controls and possible actions.

The mediation talk consists of five conversational phases: (1) framing phase, (2) topic collection, (3) working on the conflict, (4) looking for a solution (5) contract. During each phase of the mediation talk, the players are offered specific sentence openers in order to (a) provide support to the players, (b) help the chatbot to understand the general gist of each message, and (c) help the chatbot to identify the speaker. In every phase, the conversation provides the players with sentence openers indicating "affirmation", "rejection", "further inquiry", and also specific openers depending on the current phase, if needed. The players are always forced to choose one of the sentence openers, but they are able to freely finish the selected sentence.

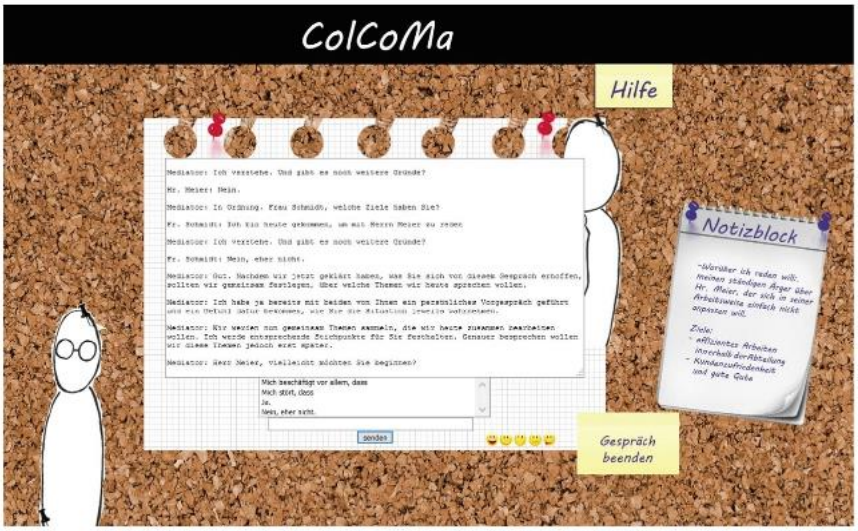

Fig. 1. Interface of the ColCoMa chat environment

The conversation can be either successfully finished in case a conflict resolution has been achieved or canceled by the mediator bot if he notices that the conversation does not advance anymore or if one of the players leaves. Irrespective of the way the mediation talk ends, the game phase is followed by a reflection phase, in which participants receive textual feedback on their overall performance during the conversation, followed by a replay session, in which the whole chat conversation is recaptured. The replay is augmented with individual feedback at certain points of interest. For example, a player will be praised for especially positive contributions or be criticized for interrupting the dialogue partner or showing inappropriate behavior. In this phase, the players are supposed to leave their role and reflect on their own behavior from an outside perspective.

4 Goals and Hypotheses

Our goal for this experiment was to investigate the question if there is a correspondence between gaze synchronicity, quality of game result (called achievement score) and the quality of the collaboration process, which can be perceived as a measurement of interactivity for this situation. How we derive these measurements is described in Sect. 6.1. for the gaze synchronicity, Sect. 6.2 for the achievement score and Sect. 6.3 for collaboration quality. We assume that the quality of collaboration as well as the successful mastering of the given task (conflict resolution) is somewhat connected to the synchronicity of the eye gaze movements. To examine the validity of this assumption, three main hypotheses were formulated:

Hypothesis 1: There is a positive relation between the convergence of visual foci of attention (gaze synchronicity) and the successful completion of the game (achievement score).
Hypothesis 2: There is a positive relation between the convergence of visual foci of attention (gaze synchronicity) and the quality of collaboration in the chat.
Hypothesis 3: There is a dynamic (time-related) congruence between similar eye gaze movements (synchronicity) and the quality of collaboration in the chat.

5 Experimental Design

Population. 20 participants (average 22.8, SD = 2.84, 5 females, 15 males) have been tested in dyads. They did not know each other prior to the study and did not meet each other in person during the experiment to avoid the possibility of any first impressions influencing their behavior during the chat situation. The participants were recruited from the FH Dortmund (university of applied sciences and arts) campus, as well as through flyers and social media.

Eye-Tracking Setup. We used two desktop-based Tobii eye-trackers, one TX300 running at 300 Hz and one X120 running at 120 Hz to track participants' gaze. For the experimental setup, the recording and later the analysis of the gaze data, the software Tobii Studio has been utilized, which offers a standard procedure for eye-tracking studies and simplifies the data analysis. An in-house server was used to synchronize and capture all eye gaze data.

Procedure. The role distribution was randomized. After the briefing and the eye-tracker calibration, the participants were introduced to the scenario in the form of a picture story, each from the perspective of the respective role. During the main game session, the participants saw the representations of the two dialog partners (the other conflict party and the mediator) and were able to communicate with each other and the mediator by means of an integrated chat.

There were two possible outcomes in this scenario: Either the participants successfully completed the conversation by achieving a conflict resolution or the mediator canceled it in case he noticed that there was no more advancement or one of the

participants left the conversation. The participants took as much time as they needed for completion (usually around 40 min). The session with the system was followed by answering several post-experiment questionnaires.

6 Method of Analysis

For the evaluation of the three main hypotheses mentioned above, we applied the following methods of analysis:

6.1 Eye-Tracking Data

In order to be able to compare the eye gaze movements of the two participants of each dyad, we defined six areas of interest. As shown in Fig. 2, the chat interface was subdivided into three different parts because otherwise the chat area would have been too extensive. Apart from the chat area, three additional areas of interest—the input area, the face of the mediator and the face of the game partner – were defined.

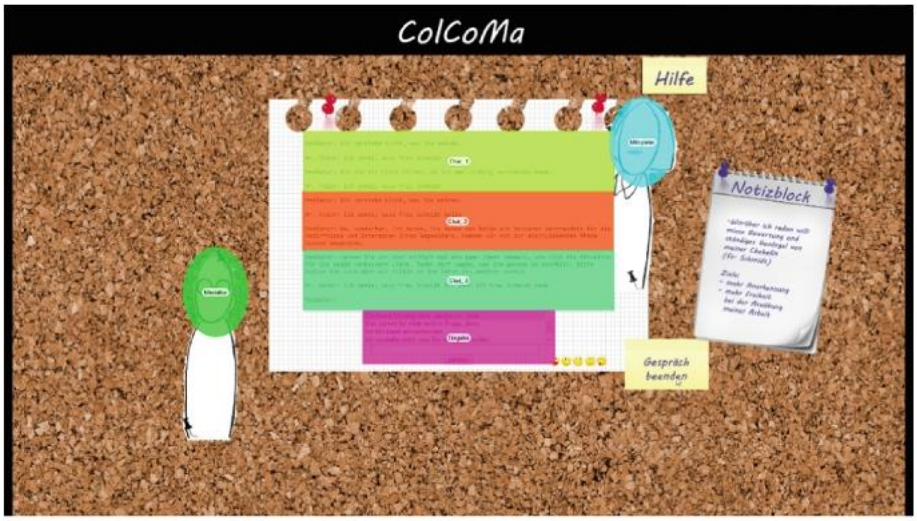

Fig. 2. Areas of Interest in the ColCoMa interface (3 sections of the chat protocol, input area, mediator face and partner face)

After the recordings, the tracking software converted the gaze movements into metric data based on these areas of interest in the form of a table. The table contains a time stamp column and six additional columns – each for one area of interest. Once a participant gazed at one of the areas of interest at a specific point of time, the respective cell was coded with a one. In contrast, a non-gazed area of interest was coded with a zero.

	A	B	C	D	E
1	LocalTimeSta	AOI[Chat_1A	AOI[Chat_ 1I	Chat1_Vergl	Chat1_Vergl
2	12:13:14	19	0	TRUE	
3	12:13:15	0	100		TRUE
4	12:13:16	0	10		
5	12:13:17	0	53		
6	12:13:18	0	35		
7	12:13:19	0	68		TRUE
8	12:13:20	23	14	TRUE	TRUE
9	12:13:21	41	0	TRUE	
10	12:13:22	52	0		

Fig. 3. Calculation of coincidences of areas of interest within 3 s windows

In order to analyze the gaze synchronicity of the participants during the game session, the recorded data needed to be compared. Due to the high amount of data, we used VBA (Visual Basic for Applications) to enable an automated comparison and evaluation of the tables. In a first step, the tables were formatted. The second step compared the tables of both participants of the same dyad. In doing so, the same area of interest was matched at a specific time stamp, using a timeframe of three seconds with one second back and one second ahead. The three second interval is an adaptation to the 2 s delay of gaze following within gaze sharing [7]. We assume a shared focus and thus synchronicity if the gaze of the two partners focused on the same area within less than 2 s difference. Since we did not (yet) use mutual gaze and thus do not have asymmetric gaze following behavior we compute the symmetric difference into the past and future second(s). A positive match for this time point (minimally one coincidence) was coded with "true". An example of applying this procedure is shown in Fig. 3.

At the end, the matches were summed up and divided through the total number of seconds to normalize against shorter or longer experiment duration. The calculations resulted in a percentage, which indicates to what extent the two participants have been looking at the same areas of interest in the same time interval during the course of the whole game. This aggregated percentage represents what we call "convergence of visual foci".

6.2 Rating Scheme for the Success in Game

To measure the success in the game we developed an *achievement score*. This achievement score reflects the participants' performance during the mediation talk based on three main criteria: Automated feedback provided by the system, successful completion of phase 2 (topic collection), and successful completion of phase 5 (contract).

The first criterion is based on the systems' automatically generated feedback. It appears at the end of the game and summarizes the participants' behavior during the game in terms of objectivity, aggressiveness and the relation between I- and You-messages. For each positive feedback, the participants receive one point in our rating, so three points in total. The second criterion refers to the second phase of the mediation talk (topic collection). In order to achieve the total of six points, the participants have to find three topics, which are needed to reach the next phase. When the process of finding these topics is not successful, the mediator interrupts the participants at some point and asks them again to think of an (further) issue they want to talk about. This process ends after three topics are found or after the mediator repeated his question three times. At this point, the mediator cancels the mediation talk when the participants still did not find the required number of topics. According to the mediator's count of repetitions, this amount is subtracted from the six possible points. The third criterion refers to the successful completion of the game. It is fulfilled when the dyad completes the game by verbally signing a contract between the parties, which includes the arrangements and rules the dyad worked out together with the mediator. For finishing the game, the participants receive three points. In total, a dyad can achieve 12 points.

6.3 Rating Scheme for Collaboration Quality

To assess collaboration quality in the chat situation we first needed to define a suitable method for analyzing the chat content of a dyad. Previous research has shown many possibilities to analyze such data, like counting certain words or analyzing the time sequences between each text message [14, 15]. However, the chat of the game ColCoMa is not suitable for any of the suggested methods due to a variety of reasons. First, this is a mediated chat in which the participants do not directly interact with each other but have a conversation guided by a mediator. Second, the structure of the chat is predefined, which makes it difficult to analyze any time related aspects of the conversation.

To find a suitable method, we first looked at all aspects of the chat conversations, which indicated a collaborative communication between the participants. We found that successful dyads had more suggestions for a solution and argued more with each other, compared to other (less successful) dyads. These findings are similar to the rating scheme of Meier et al. [11]. Their dimensions reaching consensus was about solution-focused communication, where the participants argued with each other in order to find a solution. Furthermore, it was clear that successful dyads referred more often to each other by using personal pronouns or the name of the chat partner than the less successful dyads. In addition, successful dyads talked more about their shared future or shared past with the other one. Based on these findings we developed a rating scheme to access the quality of collaboration. It includes five dimensions:

- *Argumentation:* participants discuss or bring forward arguments giving justifications (e.g. "…for this reason we are losing clients")
- *Agreement/disagreement:* participants explicitly endorse or dissent from one another (e.g. "…I agree", "…I don't agree")
- *Collaborative orientation:* participants refer to each other, ask questions, give feedback or act on topics brought up by another party ("…like Mr. Meier/Mrs. Schmidt said")

- *Solution orientation:* participants try to find a solution or make a proposal to find one ("…I'll try to pay more attention to this in future")
- *Shared awareness/reinforcing shared history:* participants share common knowledge, explain their situation ("…that my assessments of Mr. Meier's performance are always taken personally")

In order to assess the collaboration, every single chat message had to be analyzed and checked against the five dimensions. The more dimensions were matched, the higher the quality score. At the end, all matches were summed up for all chat messages in a session of a given dyad. The calculations resulted in a percentage indicating the overall collaboration quality of the chat conversation for this dyad.

In addition to the eye-tracking data and the textual chat analysis, we assessed the participants' perception of the chatbot, the perception of their partner, the performance of their partner and the collaboration. To gain this subjective data, we used several post-experimental questionnaires at the end of each game session. One of these measured human-like qualities of the mediator [16]. Additionally, we used a questionnaire to measure the interpersonal attraction [17] for the bots and the partner's perception. Furthermore, we used a questionnaire to measure the group awareness [18]. Finally, the participants were asked to rate the quality of their collaboration by a self-designed questionnaire.

7 Results

Referring to hypothesis 1, we found a highly significant correlation between the percentage value for convergence of visual foci of attention (aggregated) and the achievement score ($r = .589$, $p = .006$) as a measure for the success in the role playing game.

Relating to hypothesis 2, we found a highly significant correlation between the convergence of visual foci of attention (aggregated) and the quality of collaboration in the chat ($r = .774$, $p = .000$), which was assessed using the rating scheme presented in Sect. 6.3. There are also significant correlations between the convergence of visual foci of attention and single dimensions of collaboration quality, namely agreement/disagreement, solution orientation and shared awareness (Table 1). These findings correspond to our expectations and show that there is a connection between the synchronicity of the eye gaze movements and the performance of the participants in the game.

Table 1. Correlations between the convergence of visual foci of attention and dimensions of collaboration quality

Dimension	Gaze synchronicity
Argumentation	.228
Agreement/disagreement	.456*
Collaborative orientation	.386
Solution orientation	.509*
Shared awareness/reinforcing shared history	.609**

*$p < .05$; **$p < .01$

Surprisingly, hypothesis 3 could not be verified. There is no significant (time-related) congruence between similar eye gaze movements and the quality of collaboration in the chat. Possible reasons for this finding will be discussed in Sect. 8.

The post-questionnaires revealed interesting results, too. The participants reported above-average satisfaction with the collaboration (Fig. 3). Regarding the perspective, participants rated their own contribution better than the one of the partner or the conjoint contribution (Table 2).

Table 2. Perception of the collaboration (1 to 5 scale)

	Self		Partner		Conjoint	
	M	SD	M	SD	M	SD
Collaboration	3.65	1.13	3.50	1.05	3.55	1.19
Competence	3.50	1.05	3.30	.98	3.65	1.30

The results regarding the perception of the chatbot show that it is perceived as especially "polite" (M = 6.90 of 9), "thoughtful" (M = 6.55 of 9) and "engaging" (M = 6.45). The conversational skill of the bot was rated average to above-average (M = 5.25 of 9). Only the human-likeness of the bot was rated below-average (M = 3.95 of 9).

The perceived task attraction of the partner has been rated above-average (M = 6.90), while the social attraction achieved only an average level (M = 4.60). Furthermore, the group awareness was rated above-average.

8 Discussion

The results of the study confirmed hypothesis 1, i.e. we found a correspondence between the convergence of visual foci of attention and the success in the game. It is plausible that the extent of gaze synchronicity influences the outcome of the game and not vice versa. So we assume: the more similar the visual foci of attention, the better the collaborative performance of the two participants. Of course, other factors might have an influence on the performance, which have not been comprised in this study, so a causal relation cannot be verified.

We also found a positive relation between gaze synchronicity and the quality of collaboration (hypothesis 2), especially with the three dimensions of agreement/disagreement, solution orientation and shared awareness (on the aggregate level). Those three dimensions might be particularly good quality criteria. Here we do also suppose that a higher gaze synchronicity results in a higher collaboration quality. It could be argued that it is not necessarily the gaze synchronicity between the participants, which has a positive effect on the collaborative performance, but instead their 'general attention' (the extent to which the participants' gaze resides in the areas of interest in general). From our point of view, both the synchronicity of the eye movements and this general attention can serve as indicators for successful collaboration, and they complement one another instead of contradict.

However, hypothesis 3 could not be verified. We expected to find also a congruence between the gaze synchronicity and the quality of collaboration during the course of the game, but we found only strong correlations on the aggregate level (taking overall eye-tracking convergence as a global parameter), yet not in terms of synchronicity between convergent eye-tracking and chat interaction. We assume, that the specific nature of the chat in ColCoMa might be a reason for these results, since there are three persons involved and the two human dialog partners do not directly communicate with each other. They do always talk to the mediator and reply to his questions, and never really have the chance to react on the messages of the other conflict party immediately. Thus, the structure of the chat conversation is predefined to a certain extent and the time interval between the utterances of the different dialog partners is rather high due to the fact, that there is always the mediator writing in-between.

9 Conclusion and Future Work

The results of our study provide strong indication of that a certain convergence of the visual foci of attention between cooperation partners indicates higher success and quality of collaboration, although not all three hypotheses have been confirmed. There are quite strong correlations on the aggregate level (taking overall eye-tracking convergence as a global parameter), yet not in terms of synchronicity between convergent eye-tracking and chat interaction. This is possibly due to the specific distribution of roles in our virtual training environment. The novelty of our approach lies in the combination of this eye-tracking method with virtual role play, especially considering a scenario with two human players and a chatbot.

Due to the relatively small amount of participants, no generalization should be made here, but the results are promising and should be elaborated on in larger studies. Further studies should be conducted to re-check synchronicity between convergent eye-tracking and collaboration quality with a modified chat environment which allows direct communication between the cooperation partners.

Future work could also include chat awareness tools, which e.g. indicate if a chat message has been read or even visualize the eye gaze of the chat partner. Recent studies indicate that real-time mutual gaze perception via awareness visualisations enhances collaborative learning and collaboration quality [19]. Finally, further studies are needed to identify more factors, which are influencing the quality of collaboration in order to improve collaborative processes.

References

1. Whitton, N., Hollins, P.: Collaborative virtual gaming worlds in higher education. ALT-J, Res. Learn. Technol. **16**(3), 221–229 (2008)
2. Greco, M.: The use of role-playing in learning. In: Connolly, T., Stansfield, M., Boyle, L. (eds.) Games-Based Learning Advancements for Multi-Sensory Human Computer Interfaces, pp. 158–173. Information Science Reference, New York (2009)

3. Slator, B.M., Chaput, H.: Learning by learning roles: a virtual role-playing environment for tutoring. In: Frasson, C., Gauthier, G., Lesgold, A. (eds.) ITS 1996. LNCS, vol. 1086, pp. 668–676. Springer, Heidelberg (1996). https://doi.org/10.1007/3-540-61327-7_167
4. Dillenbourg, P.: What do you mean by collaborative learning? In: Dillenbourg, P. (ed.) Collaborative-Learning: Cognitive and Computational Approaches, pp. 1–19. Elsevier, Oxford (1999)
5. Dillenbourg, P., Baker, M., Blaye, A., O'Malley, C.: The evolution of research on collaborative learning. In: Reimann, P., Spada, H. (eds.) Learning in Humans and Machine: Towards an Interdisciplinary Learning Science, pp. 189–211. Elsevier, Oxford (1996)
6. Schneider, B., Pea, R.: Real-time mutual gaze perception enhances collaborative learning and collaboration quality. Int. J. Comput.-Support. Collab. Learn. **8**(4), 375–397 (2013)
7. Richardson, D.C., Dale, R.: Looking to understand: the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. Cogn. Sci. **29**(6), 1045–1060 (2005)
8. Cherubini, M., Nuessli, M., Dillenbourg, P.: Deixis ad gaze in collaborative work at a distance (over a shared map): a computational model to detect misunderstandings. In: Proceedings of the 2008 Symposium on Eye Tracking Research & Applications, pp. 173–180. ACM, New York (2008)
9. Jermann, P., Mullins, D., Nuessli, M.-A., Dillenbourg, P.: Collaborative gaze footprints: correlates of interaction quality. In: CSCL 2011 Conference Proceedings, Long Papers, vol. 1, pp. 184–191. International Society of the Learning Science, Hong Kong (2011)
10. Schlösser, C., Schlieker-Steens, P., Kienle, A., Harrer, A.: Using real-time gaze based awareness methods to enhance collaboration. In: Baloian, N., Zorian, Y., Taslakian, P., Shoukouryan, S. (eds.) CRIWG 2015. LNCS, vol. 9334, pp. 19–27. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22747-4_2
11. Meier, A., Spada, H., Rummel, N.: A rating scheme for assessing the quality of computer-supported collaboration processes. Int. J. Comput.-Support. Collab. Learn. **2**(1), 63–86 (2007)
12. Schneider, B., Pae, R.: Toward collaboration sensing. Int. J. Comput.-Support. Collab. Learn. **9**(4), 371–395 (2014)
13. Sharma, K., Chavez-Demoulin, V., Dillenbourg, P.: An application of extreme value theory to learning analytics: predicting collaboration outcome from eye-tracking data. J. Learn. Anal. **4**(3), 140–164 (2017)
14. Bratitsis, T., Dimitracopoulou, A., Martínez-Monés, A., Marcos, J.A., Dimitriadis, Y.: Supporting members of a learning community using Interaction Analysis tools: the example of the Kaleidoscope NoE scientific network. In: 8th IEEE International Conference on Advanced Learning Technologies (ICALT 2008). IEEE Computer Society, Santander (2008)
15. Kahrimanis, G., Chounta, I.-A., Avouris, N.: Study of correlations between logfile based metrics of interaction and the quality of synchronous collaboration. Int. Reports Socio-Inform. **7**(1), 24 (2010)
16. Holtgraves, T.M., Ross, S.J., Weywadt, C.R., Han, T.L.: Perceiving artificial social agents. Comput. Hum. Behav. **23**, 2163–2174 (2007)
17. McCroskey, J.C., McCain, T.A.: The measurement of interpersonal attraction. Speech Monogr. **41**, 261–266 (1974)
18. Mock, A.: Open (ed) classroom–who cares? MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung **28**, 57–65 (2017)
19. Schlösser, C., Cedli, L., Schröder, B., Kienle, A.: Beyond gaze cursor - exploring information-based gaze sharing in chat. In: Workshop on Communication by Gaze Interaction, COGAIN 2018, Warsaw, Poland. ACM, New York (2018)

4 CASE STUDY: CUSTOMER COMPLAINT MANAGEMENT (1)

SUPPORTING GROUP REFLECTION IN A VIRTUAL ROLE-PLAYING ENVIRONMENT

This paper was presented at the 8th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN) 2016 in the context of a workshop titled "Virtual Agents for Social Skills Training" and published as part of the Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering book series (LNICST, volume 178). While the conference itself covered the topics of serious games, interaction technologies, persuasion and motivation, exertion games, game studies, and novel applications and tools, the workshop explicitly focused on virtual agent applications that are intended for the training of social skills.

Reprinted by permission from Springer Nature: Springer Nature, Intelligent Technologies for Interactive Entertainment. INTETAIN 2016. Othlinghaus, J. & Hoppe, H. U.: Supporting Group Reflection in a Virtual Role-Playing Environment. © ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (2017)

Author	**Contribution**	**%**
Julia Othlinghaus-Wulhorst	- Main responsibility in the design, implementation and evaluation of the game - Writing of main manuscript	80%
H. Ulrich Hoppe	- Research planning and guidance - Text revision	20%

Supporting Group Reflection in a Virtual Role-Playing Environment

Julia Othlinghaus(✉) and H. Ulrich Hoppe

Faculty of Engineering, Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen, Lotharstr. 63/65, 47048 Duisburg, Germany
{othlinghaus,hoppe}@collide.info

Abstract. This paper presents an approach to supporting group reflection in a virtual role-playing environment with intelligent support designed for the training customer complaint management in electronic shops. The single-player design involves a player and an AIML chat bot in a 2D web-based virtual environment. Building on this, a group reflection tool was designed, which is supposed be used in a training center environment. It features a dashboard design which includes different visualizations of player performance based on automated individual analyses of players' communicative behavior, as well as enriched replays of their conversations, and the ability to make annotations. The separation of the application into the actual role-playing game and the group reflection tool is assumed to support the learning process of responding to customer complaints by changing perspective, receiving feedback, and recognizing different ways of problem solving.

Keywords: Group reflection · Role-play · Intelligent support · Serious games · Multi-agent architecture · Chat bots

1 Introduction

Role-play allows participants to "play a role" in a situation: to act as themselves or otherwise in an environment without fearing irreversible consequence [1]. This makes it essential for a wide-range of education and training scenarios [2]. Authentic simulated environments provide learners with meaningful and near-real experiences: they "learn by doing" [3]. Customer complaint handling skills are often trained by role-playing with simulated customers. Correctly handling customer complaints has become an increasingly important social skill [4, 5]. However, traditional role-play in this context can be time-consuming, hard to administer, and lacking repeatability, while virtual role-plays can provide portable, safe and continuable environments.

This paper presents an attempt to train customer complaint handling skills by building on role-playing in a virtual simulation environment and providing a tool for after-action review. This takes the form of guided group reflection based on automated analysis of player performances.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2017
R. Poppe et al. (Eds.): INTETAIN 2016, LNICST 178, pp. 292–298, 2017.
DOI: 10.1007/978-3-319-49616-0_30

2 Background and Related Work

Serious Role-Playing Games. Serious games increase in being acknowledged as efficient and powerful tools for promoting learning and encouraging behavioral change [6], and thus have a great potential for professional training [7]. This article is in line with the work of Malzahn et al. [8], Emmerich et al. [9] and Ziebarth et al. [10]. It focuses on serious role-playing games in 2D and 3D environments for training specific social skills and follows a scenario-based approach. In scenario-based learning environments, conditions, characters, circumstances and parameters are drawn to simulate a real-life context for learning [11]. Table 1 provides a summary of the existing approaches, the training scenarios and their key aspects. The distinctive feature of the approach presented here is the explicit support for group reflection.

Table 1. Overview of existing approaches.

Year	Application domain	Constellation	Support/emphasis
2010	Job interviews	Single-user + chat bot	Scaffolding, Evaluation
2012	Conflict management	Multi-user + chat bot	Collaboration
2014	Patient-centered medical interviews	Single-user + chat bot	Gamification
2016	Customer complaint management	Single-user + chat bot	Explicit group reflection support

Customer Complaint Management. The successful handling of customer problems enhances customer satisfaction, trust and commitment [4], which are essential elements in establishing strong long-term customer-firm relationships [5, 12] and building sustainable market share [13]. While increasing investments in handling complaints can be recognized, firms are lacking effective strategies and programs [4]. Principal evaluative criteria of customer complaints are: (1) the resolution procedures, (2) the interpersonal communications and behavior, and (3) the outcome [4]. General guidelines for handling complaints are, amongst others, provided by the British *Institute of Customer Service* [14]. They include concrete recommendations on how to behave towards a complaining customer, such as thanking the customer for complaining, putting oneself in the place of the customer, always assuming that the customer has a valid point, getting all the facts, correcting the mistake, and responding at any time. Those guidelines have served as a compendium for the performance evaluation and rating.

Group Reflection. Reflection is an important activity in which people recapture, rethink and evaluate their experience in order to lead to new understandings and appreciations, which is very important for learning [15]. Learning processes cannot exclusively be reflected by oneself, but preferably in groups, which enables collective exchange and thus collaborative learning [16]. The term 'group reflection' describes a sort of meta-communication within a group about the learning process [17]. Furthermore, Kim et al. found out that effective instructor intervention is a crucial component leading to the better performance of a group in terms of learning [18].

3 *CuCoMaG*: The Game

The effects of a role-play, in particular the pedagogical outcome, usually depends on the post-role-play reflection, since without feedback the transfer to real-world situations cannot be secured [2]. After-action review is a method that helps learners to identify and share effective practices and strategies derived from the experience [19]. This requires a change of perspective. In our approach, this is facilitated by the differentiation of phases of immersion (role-play) and reflection (group session), which is considered to be conducive to learning as meta-cognitive activities are advanced [8].

Game Design. In the role-playing part, the player has the role of an employee who is responsible for customer service in a shoe-selling online shop. He finds himself in a conversation with a chat bot acting as a customer, who reports a certain problem. The player communicates with the customer through a simple chat environment (Fig. 1). Each chat message consists of a sentence opener which the player needs to select from a predefined set, and free text, which (a) gives the player the possibility to express himself more naturally and (b) allows a more detailed analysis of the communicative behavior. Based on the sentence openers, the chat bot can interpret the player's intentions and is able to detect and adequately react to e.g. aggressive or rude behavior by the player.

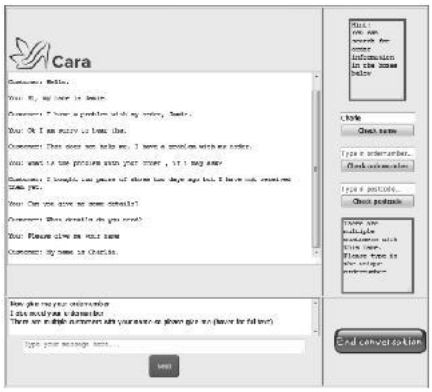

Fig. 1. Chat interface

Each scenario is kept simple and comprehensible to support understanding of the role and task, and to allow the player to focus on the problem-solving process. The main goal is to come to a resolution at the end of the conversation by showing appropriate behavior and choosing beneficial strategies of managing the complaint. A scoring system was implemented to evaluate the individual communicative behavior of each player and to make the performances of different players comparable. Relevant factors that influence the player's score in either a positive or a negative way are politeness, aggressiveness, rudeness, use of forbidden terms and phrases, message time,

message length, moments of silence, answer quality (unhelpful/neutral/helpful), the total amount of answers and the quality of the final solution.

Implementation. The conversational behavior of the chat bot has been implemented using the *Artificial Intelligence Markup Language* (AIML), an XML-based solution for passive chat bots, which follows a simple pattern-matching mechanism [20]. The passive nature and the limited capabilities of AIML have required creative work-arounds in order to enable the bot to become active (by using external triggers) and to show appropriate reactions to the player's input (by preprocessing and using sentence openers) without having an extremely complex script.

In order to ensure platform independence and easy access, the logic and interface of the game client has been realized as a web-based application using common web technologies, such as HTML, CSS and JavaScript. The analysis of the player's performance has been designed as a multi-agent system which includes 11 agents (individual programs implemented in C#) in addition to the client. The multi-agent architecture has resulted in a loosely coupled system that can be easily extended and adapted. It is based on *SQLSpaces*, an implementation of the tuple space concept, which supports various programming languages and is built on a relational database [21]. Each of the agents is responsible for one certain aspect of either input analysis or game control. They mainly check the text input against predefined lists of words, expressions or phrases, or measure certain quantitative aspects, such as the time needed to send the message. The results of the agents' evaluation influence the player's score as well as the answering behavior of the chat bot.

4 Group Reflection Support

The reflection tool has been designed to be used in training centers and requires a trainer or expert to support the group reflection process. This involves a group of participants who have played the game before the start of the session. Since the tool has been realized as a detached application, the reflection phase does not necessarily take place right after the gaming sessions, so the trainer has time to inspect the material and to prepare the group discussion. The tool provides comprehensive preparation and visualization of the analyses' results to the trainer to support the group reflection process.

The design of the reflection tool has been kept simple and plain (Fig. 2). The application allows the trainer to load the game data of different participants; it provides a transcript of their chat conversations annotated with the analyses' results; it offers different chart-based visualizations of each user's score and certain aspects of the communicative behavior; it also facilitates the option to make notes with the help of a notepad. The transcripts give the trainer the possibility to replay the whole chat conversation of each participant or just certain passages from them. The transcripts are enriched with annotation tags showing the analysis agents' findings. In regards to the different chart-based visualizations, the trainer can select one or more players and one or more evaluative factors, such as politeness, aggression, rudeness, moments of silence, answer quality (unhelpful, neutral and helpful) and no-go answers, to display

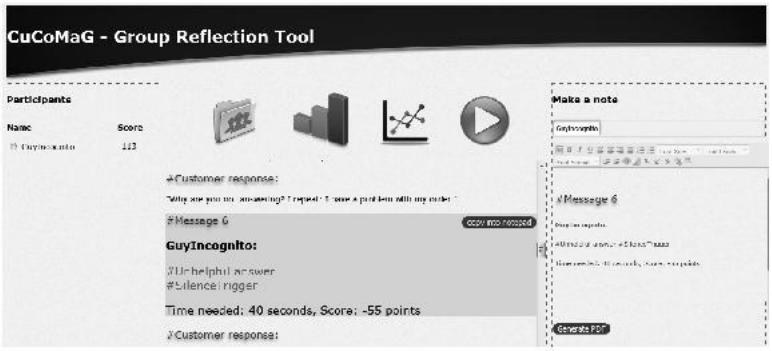

Fig. 2. Group reflection tool

them in the form of a bar chart or line chart. The bar chart shows how many times the selected factors occurred in the selected participants' performances, while the line chart presents the development of the participants' scores over time, with the option to highlight occurrences of the selected factors. The line chart is directly connected to the chat transcripts. This means that the trainer can switch from the chart to a certain point of the conversation by clicking on the graph. The notepad features the option to make notes for each participant. It is also possible to copy parts of the transcripts into the notepad and annotate them. The content of the notepad can be exported at the end of the reflection session to create a report for each participant.

The reflection tool has been designed as a web-based application and implemented using HTML, CSS, JavaScript, as well as JQuery[1] and Highcharts[2] libraries.

5 Conclusion and Future Work

We have presented a virtual role-playing environment for the training of customer complaint handling with group reflection support. The unique feature of this approach is the combination of an immersive role-playing scenario supported by an AI-controlled chat bot, with a separate group reflection phase reinforced by an evaluation tool based on automated performance analysis. Using a chat-based virtual role-play for training customer complaint handling in electronic shops is especially useful, since it provides a realistic training environment which simulates everyday work situations. The group reflection session is supposed to be guided by a trainer, who can use the tool to arrange the interactive after-action review process. With the help of the tool, the trainer can show important sequences from the participants' chat conversations, review specific actions and reasons for the outcome of the game, highlight certain aspects of the communicative behavior, give feedback, and initiate group

[1] http://www.jquery.com/.
[2] http://www.highcharts.com/.

discussions in order to enable reflection by the participants on their actions and to help them improve their performance in the future.

Due to the flexible multi-agent system architecture, the application can easily be adapted to other contexts and scenarios. Future tasks include usability and field studies, as well as the generation of additional scenarios featuring different customer and problem types to increase repeatability and diversity.

References

1. Ladousse, G.P.: Role Play. Oxford University Press, Oxford (1987)
2. Lim, M.Y., Aylett, R., Enz, S., Kriegel, M., Vannini, N., Hall, L., Jones, S.: Towards intelligent computer assisted educational role-play. In: Chang, M., Kuo, R., Kinshuk, Chen, G.-D., Hirose, M. (eds.) Edutainment 2009, LNCS, vol. 5670, pp. 208–219. Springer, Heidelberg (2009)
3. Slator, B.M., Chaput, H.C.: Learning by learning roles: a virtual role-playing environment for tutoring. In: Frasson, C., Gauthier, G., Lesgold, A. (eds.) ITS 1996. LNCS, vol. 1086, pp. 668–676. Springer, Heidelberg (1996). doi:10.1007/3-540-61327-7_167
4. Tax, S.S., Brown, S.W., Chandrashekaran, M.: Customer evaluations of service complaint experiences: implications for relationship marketing. J. Mark. **60**, 60–76 (1998)
5. Sirdeshmukh, D., Singh, J., Sabol, B.: Consumer Trust, Value, and Loyalty in Relational Exchanges. J. Mark. **66**(1), 15–37 (2002)
6. Cheong, Y.-G., Khaled, R., Grappiolo, C., Campos, J., Martinho, C., Ingram, G., Paiva, A., Yannakakis, G.N.: A computational approach towards conflict resolution for serious games. In: Proceedings of the Sixth International Conference on the Foundations of Digital Games. ACM, New York (2011)
7. Marr, A.C.: Serious Games für die Informations- und Wissensvermittlung: Bibliotheken auf neuen Wegen. Dinges & Frick, Wiesbaden (2010)
8. Malzahn, N., Buhmes H., Ziebarth S., Hoppe H.U.: Supporting reflection in an immersive 3D learning environment based on role-play. In: EC-TEL 2010, pp. 542–547. ACM, New York (2010)
9. Emmerich, K., Neuwald, K., Othlinghaus, J., Ziebarth, S., Hoppe, H.U.: Training conflict management in a collaborative virtual environment. In: Herskovic, V., Hoppe, H.U., Jansen, M., Ziegler, J. (eds.) CRIWG 2012. LNCS, vol. 7493, pp. 17–32. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33284-5_2
10. Ziebarth, S., Kizina, A., Hoppe, H.U., Dini, L.: A serious game for training patient-centered medical interviews. In: IEEE 14th International Conference on Advanced Learning Technologies (ICALT), pp. 213–217. IEEE, New York (2014)
11. Herrington, J., Oliver, R., Reeves, T.C.: Patterns of engagement in authentic online learning environments. In: Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2001. AACE, Chesapeake (2002)
12. Orshinger, C., Valentini, S., de Angelis, M.: A meta-analysis of satisfaction with complaint handling in services. J. Acad. Mark. Sci. **38**(2), 169–186 (2010)
13. Urban, G.L., Sultan, F., Qualls, W.J.: Placing trust at the center of your internet strategy. Sloan Manag. Rev. **42**, 39–49 (2000)
14. The Institute of Customer Service. https://www.instituteofcustomerservice.com/research-insight/guidance-notes/article/handling-complaints

15. Boud, D., Keogh, R., Walker, D.: Promoting reflection in learning: a model. In: Boud, D., Keogh, R., Walker, D. (eds.) Reflection: Turning Experience into Learning, pp. 18–40. Kogan Page, London (1985)
16. Schuster, R.J.: Gruppenreflexion als Kommunikationsinstrument. In: Management und Wirtschaft. Schriftenreihe zur wirtschaftswissenschaftlichen Forschung, vol. 13, pp. 7–23. Fachhochschule des bfi Wien, Wien (2010)
17. Hilzensauer, W.: Theoretische Zugänge und Methoden zur Reflexion des Lernens. Ein Diskussionsbeitrag. In: Häcker, T.H., Hilzensauer, W., Reinmann, G. (eds.) Bildungsforschung, vol. 5(2) (2008)
18. Kim, P., Hong, J.-S., Bonk, C., Lim, G.: Effects on group reflection variations in project-based learning integrated in a Web 2.0 learning space. In: Interactive Learning Environments, vol. 19(4), pp. 333–349. Routledge, London (2011)
19. Morrison, J.E., Meliza, L.L.: Foundations of the after action review process. In: ARI Special Report, vol. 42. U.S. Army Research Institute for the Behavioral and Social Sciences, Arlington (1999)
20. Wallace, R.: The Elements of AIML Style. ALICE AI Foundation (2004)
21. Weinbrenner, S.: SQLSpaces – A Platform for Flexible, Language-Heterogeneous Multi-Agents Systems. University of Duisburg-Essen (2012)

5 CASE STUDY: CUSTOMER COMPLAINT MANAGEMENT (2)

TRAINING CUSTOMER COMPLAINT MANAGEMENT IN A VIRTUAL ROLE-PLAYING GAME: A USER STUDY

This paper was presented at the 14th European Conference on Technology Enhanced Learning (EC-TEL) 2019 and published as part of the Lecture Note in Computer Science book series (LNCS, volume 11722). 41 research papers have been selected from 149 submissions (27.52% acceptance rate). The contributions reflect the debate around the role of and challenges for meaningful technologies and advances such as artificial intelligence and robots, augmented reality and ubiquitous computing technologies and connecting them to different pedagogical approaches, types of learning settings and application domains that can benefit from these technologies.

Reprinted by permission from Springer Nature: Springer Nature, Transforming Learning with Meaningful Technologies. EC-TEL 2019. Othlinghaus-Wulhorst, J., Mainz, A. & Hoppe, H. U.: Training Customer Complaint Management in a Virtual Role-Playing Game: A User Study. © Springer Nature Switzerland AG (2019)

Author	**Contribution**	**%**
Julia Othlinghaus-Wulhorst	- Main responsibility in the design, implementation and evaluation of the game - Conceptualization of the evaluation study - Writing of main manuscript	60%
Anne Mainz	- Re-design of the prototype - Conduction of the evaluation study - Statistical Evaluation	30%
H. Ulrich Hoppe	- Research planning and guidance - Text revision	10%

Training Customer Complaint Management in a Virtual Role-Playing Game: A User Study

Julia Othlinghaus-Wulhorst(✉), Anne Mainz, and H. Ulrich Hoppe

University of Duisburg-Essen, Duisburg, Germany
{othlinghaus, hoppe}@collide.info,
anne.mainz@stud.uni-due.de

Abstract. Handling customer complaints properly, especially through chat or phone-based interaction, has become an increasingly important social skill and is the subject of professional training in companies, markets, and multinational corporations. In order to develop such skills, training methods can involve videos and role plays. Virtual role play scenarios can provide a fairly authentic experience of realistic conflict situations with customers and allow for trying out different problem-solving strategies without consequences in the real world. This paper presents an attempt to train customer complaint handling through an educational role-playing game based on theories of consumer psychology and complaint management using a chatbot system with intelligent support. The playability, game experience, and perception of the virtual role play environment, as well as the interaction with the chatbot, have been evaluated in a mixed method study. The results indicate that the idea and approach of the game, in general, are assessed positively and the scenarios are perceived as useful and realistic. Furthermore, the study confirms that the chatbot's conversation style is influencing the game experience and the perception of the chatbot significantly.

Keywords: Virtual role play · Intelligent support · Customer complaint management

1 Introduction

Especially in the e-business sector, customers can choose between a variety of different products and providers, which makes customer loyalty a core challenge. Excellent online customer service is one of the most important factors for ensuring customer satisfaction [1]. Accordingly, handling customer complaints has been identified as an important social skill and is the subject of professional training in companies, markets, and multinational corporations [2]. Video-based learning and role plays can be utilized in this context to consolidate the proper concepts and to develop professional behavior when handling customer complaints [3]. Especially virtual role plays help people experience conflict situations with customers and learn how to handle complaints. They provide a safe environment for trying different problem-solving strategies, and although their actions have no consequences in the real world, this training can prepare them to react adequately in similar situations.

© Springer Nature Switzerland AG 2019
M. Scheffel et al. (Eds.): EC-TEL 2019, LNCS 11722, pp. 436–449, 2019.
https://doi.org/10.1007/978-3-030-29736-7_33

This paper presents an attempt to train customer complaint handling through an educational role-playing game based on theories of customer psychology and complaint management using chatbots and intelligent support. At this point, the main evaluation questions address playability, usability, and perceived authenticity of the environment. A user study has been conducted to investigate these aspects. In the rest of this paper, we first elaborate on the background before presenting the study and its results.

2 Background and Related Work

The following subsections provide an overview of the two main fields of interest that constitute the basis for the application design: educational games for the training of soft skills and customer complaint management as the field of application.

2.1 Educational Games for Soft Skills Training

Digital role-playing and simulation-based training systems have been increasingly adopted for the training and development of soft skills [4]. The so-called technology enhanced educational role-playing games (EduTechRPGs) are digital environments that support the training of soft skills through the application of psycho-pedagogical methodology. The term soft skills describes personal attributes or traits that express how people know and manage themselves and their relationships with others [4]. It is a broad concept that includes many dimensions of the personal sphere development involving emotional, behavioral, and cognitive components [4]. The goal of EduTechRPGs is to combine education and fun, thereby increasing the (intrinsic) motivation of the players. One major advantage of EduTechRPGs is their ability to promote learning by doing. Players are supposed to undergo an active learning process of experience and reflection, which imparts soft skills in the best possible way.

There is a number of EduTechRPGs addressing different social skills. ENACT (*Enhancing Negotiation skills through online Assessment of Competencies and interactive mobile Training*) [4] is a 3D single-player game to assess and train a user's negotiation and communication skills. In the game, two on-stage agents represented by 3D avatars simulate a dialog between two people. One is controlled by the player and the other by a AI-controlled bot. The simulation includes three dimensions of communication: verbal (the words used in a sentence), para-verbal (tone, pitch, and volume of the voice), and non-verbal (body language), and has eight different scenarios. An important aspect of the game is the assessment element, which allows the measuring of soft skills based on a psychometric approach.

Virtual Leader [5] is a role-playing based simulation program, which is supposed to help students practice different leadership styles and approaches in a 3D environment. In the game, players participate in virtual business meetings with animated computer-controlled characters. The game includes five scenarios with increasing complexity. It was designed to provide an immersive environment to practice leadership skills like negotiation, collaboration, influencing, and conflict resolution, and provides immediate feedback in the form of a leadership score that is based on their effectiveness in achieving specific scenario goals.

ColCoMa [6] is a collaborative game for training workplace-oriented conflict management in a role-playing scenario. It employs two human actors in the role of the conflicting parties and an AI-controlled chatbot in the role of a mediator, who is moderating a mediation talk. The main goal of the players is to resolve the conflict by showing appropriate and constructive behavior during the conversation. The learning process is supported by adaptive feedback based on an individual performance analysis. The idea behind this is that players experience enhanced self-understanding and immersion through collaborative play, which is expected to effectively foster their conflict resolution skills.

Ziebarth et al. [7] developed a web-based game for medical students to support the training of patient-centered medical interviews. Here, players assume the role of a locum doctor for family medicine, and their goal is to find out as many of the patient's symptoms as possible within a given time frame. To identify a symptom, the player has to communicate with the patient via text input and non-verbal actions. The behavior of the patient depends on the level of trust and empathy the players have established during the conversation. Post-role-play reflection is supported by a recording of the role play session, which is further enhanced by the results of an automated analysis of the communication behavior based on models of doctor-patient communication (and general communication) used to describe general rules and strategies for medical interviews.

While all of the above games address the training of specific social skills, there is no existing approach explicitly targeting the training of customer complaint management strategies. Also, most of the studies do not come with thorough empirical evaluation. Our work aims to create a meaningful and structured approach for training customer complaint management using role play based on best practices of complaint management.

2.2 Customer Complaint Management

Complaint management can be understood as the complete system provided by the company that affords the opportunity to resolve complaints [8]. Original complaint channels, such as telephone, mail, or even personal conversations, have been more and more replaced by electronic channels (for example, email, social media, or specially created complaint platforms) [1]. The resolution of a complaint is always associated with costs: employees have to be hired, compensations for customers have to be made (e.g., refunds, repair service), and much more. Nevertheless, the mathematical model of Fornell and Wernerfelt [8] suggests that companies should encourage customers to complain and compensate them generously because complaint management serves as an effective tool for customer retention by increasing the expected benefits of the purchase for the customer. Even if the complaints are objectively not justified, it can make economic sense to react fairly, as in most cases complaints are considered to be justified from the customers' perspective [9].

There are three groups of measures available as basic solution possibilities to customer problems [9]: financial, tangible, and intangible. *Financial* solutions include money return, price reduction, and compensation for damages. *Tangible* solutions are payments in kind like exchange, repair, another product or gift. *Intangible* reactions

include all customer-oriented forms of communication that aim to reduce the customer's dissatisfaction, such as information, explanation, and apology. The choice of the appropriate compensation is confined by product-specific factors and cost considerations. According to Chase and Dasu [10], financial or tangible solutions are appropriate in case of production mistakes, whereas intangible solutions are advisable in case of corporate malpractice.

Stauss [11] differentiates between two dimensions of complaint satisfaction: outcome complaint satisfaction and process complaint satisfaction. *Outcome complaint satisfaction* encompasses the evaluation of what the customer actually gets from the company as a compensation, while *process complaint satisfaction* refers to the evaluation of how the complaint is handled. Factors creating process complaint satisfaction are access, friendliness, empathy, individual handling, effort, active feedback, reliability, and speed of response [11].

3 Virtual Role Play Environment: CuCoMaG

Based on the idea to implement a virtual customer complaint management training embedded in a role-playing scenario, we designed the EduTechRPG *CuCoMaG (Customer Complaint Management Group reflection)*. In this game, the player assumes the role of a customer service employee in *LittleOnes*, a fictitious company producing and selling personalized clothing for children via an online shop. Complaint management is particularly important for such a company because it sells sensory products, has a large number of competitors, and high quality elasticity is possible [8]. However, children's clothing does not require complex warranty regulations. In addition, clothing and accessories are the product categories most commonly associated with complaints [1].

3.1 Game Design

Players find themselves in a conversation with a chatbot in the role of a complaining customer, who has a specific problem. The player communicates with the customer through a simple chat interface (Fig. 1). In order to create a chat message, the player has to select a sentence opener from a predefined set and supplement it with free text. The sentence openers (a) provide support to the player and (b) help the chatbot to understand the intention and general gist of each message. The free text supplementation enables players to express themselves more naturally and also allows a more detailed evaluation of a player's communicative behavior. The offered set of sentence openers depends on the phase of the conversation. The player can also access the company's database to search for additional information on a customer and their order, which is necessary in order to receive all information required to resolve the situation. In summary, the user choices are: (1) selection of a predefined sentence opener, (2) input of free text to complete the user message, (3) information retrieval using the database.

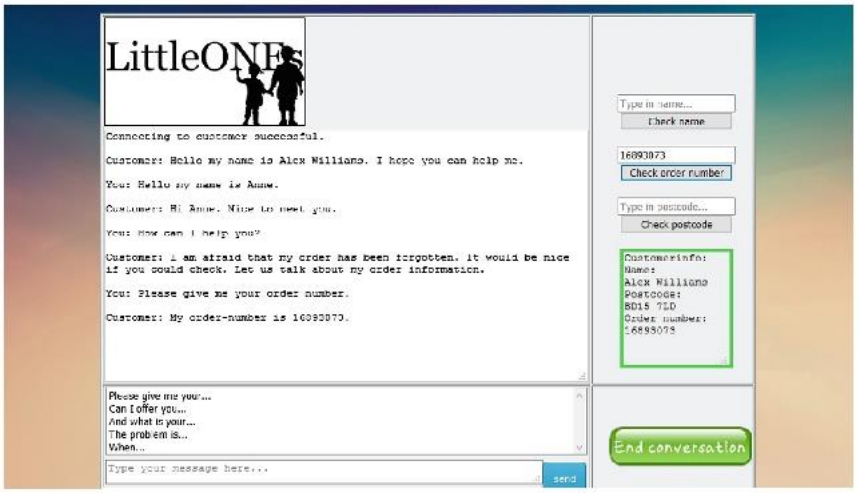

Fig. 1. Chat interface of CuCoMaG

The game includes three different scenarios. The scenarios differ based on the type of customer used, especially in terms of (a) conversation style based on the model of Rahim and Bonoma [12], who differentiated between five different styles of handling interpersonal conflicts, and (b) the problem situation of the customer based on the findings of a study conducted by Cho et al. [1], who investigated current sources and causes of online complaints, and thus (c) in the level of difficulty. Each scenario consists of five conversational phases following the typical structure of a complaint conversation according to Stauss and Seidel [9]: (1) greeting phase, (2) aggression-reduction phase, (3) conflict-settlement phase, (4) problem-solution phase, and (5) conclusive phase.

The first scenario serves as an introductory level including a tutorial. The customer in this scenario can be classified as an *integrating* customer, who is open to reach a solution acceptable for both parties and exhibits problem-solving behavior. The customer's problem in this scenario is the third most common cause of non-public online customer complaints [1]: delivery problems. The aim of this scenario is to help the player becoming acquainted with the user interface and experience the basic milestones of the complaint conversation.

In the second scenario, the level of difficulty increases. The customer is emotional about the problem and must be calmed down. According to the classification of Rahim and Bonoma [12], this customer is considered a *compromising* customer. The customer's problem is the most common problem within non-public online complaints [1]: he has, among other things, problems with the customer service. This also makes the customer a *follow-up complainant*, as it is the second time that he has contacted the customer service about the same problem [9]. The goal of the scenario is to pass through all five phases of a complaint process successfully.

The third scenario is the one with the highest level of difficulty. The customer in this scenario can be classified as a *dominating* customer [12] and a grouser [9].This type

of customer tries to force a solution that is optimal for them and is looking for a continuation of the conflict, while showing little or no understanding for the other side. The customer has problems with the business terms and conditions, which is the second most common problem with non-public online complaints according to Cho et al. [1]. The costumer is not reasonable and reacts abusive. The player's best result may be not responding to the customer's provocations and eventually ending the conversation. This is called *active farewell* [9]. The goal of this scenario is to deal with extreme situations and to prove the player's ability to deal with provocations and difficult customers.

Each scenario has three possible outcomes: (1) The player reaches a predefined end state of the conversation, (2) the player leaves the conversation, (3) the player does not reply for a certain amount of time and fails to react to repeated requests of the customer to answer so the customer terminates the conversation and leaves the chat.

To increase the learning effect of this virtual role play, it is followed by a group reflection phase based on an automated analysis of player performances. Reflection, and group reflection in particular, is a successful tool to improve learning processes [13]. A tool designed for supporting the group reflection phase visualizes the analysis of data generated from the individual player's behaviors in a dashboard design. It is assumed that the separation into the actual role-playing game (immersive phase) and the group reflection session (reflective phase) supports the learning process [6, 7]. It is important to note that the group reflection phase has not been part of our study.

3.2 Implementation

The conversational logic of the customer chatbot has been implemented using the *Artificial Intelligence Markup language* (AIML). AIML is an XML-based solution for intelligent chatbots [14]. The flow of the dialog was first specified in UML activity diagrams and later transcribed into AIML scripts using the *GaitoBot*[1] AIML editor. The limited capabilities and the passive nature of AIML required several creative workarounds: (1) improve the appropriateness of the chatbot's responses to player input by preprocessing and using sentence openers, (2) use external triggers to enable the bot to become active when needed, (3) use atomic patterns [15] to reduce possible text inputs to their semantic content in order to create maximally efficient scripts, and (4) use variables to control the flow of the conversation and to enable the chatbot to "remember" past in- and outputs despite the simple stimulus response structure of the AIML scripts.

The logic and interface of the game client have been designed as a web-based application using common web technologies such as HTML, CSS, and JavaScript to ensure easy access and platform independency. The backend of the game has been implemented as a multi-agent blackboard system and consists of twelve program modules (agents) that are running independently from each other. The information exchange between the agents and the client is established by the use of an implementation of the TupleSpace concept called *SQLSpaces* [16]. According to the blackboard paradigm, the client and all agents only communicate with the central blackboard (and not one-to-one), writing and reading tuples in order to exchange information. As a result, agents can

[1] www.gaitobot.de.

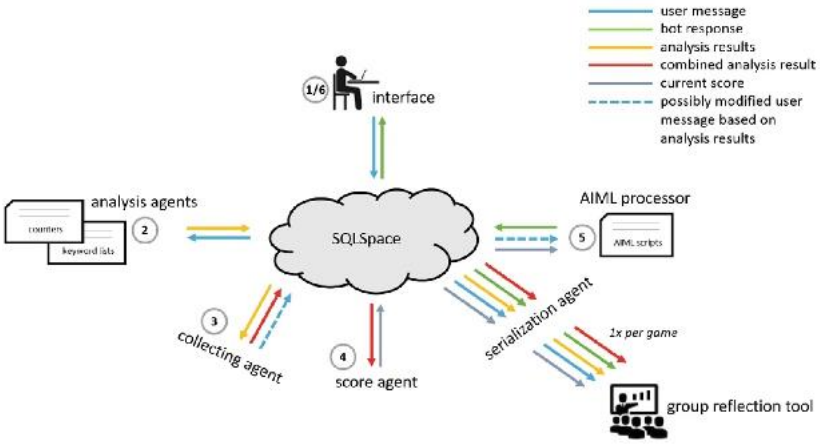

Fig. 2. Architecture and message flow of the multi-agent system

easily be added, amended, or replaced if necessary, which characterizes this loosely coupled and adaptive system. All agents are implemented in Java.

The overall process is displayed in Fig. 2. The *user interface* sends each user message to the tuple space where it is processed in parallel by several *analysis agents*. Each of them is responsible for one certain aspect of input analysis. Their main task is to check the input against predefined lists of keywords, expressions, or phrases (e.g. to find inappropriate, rude, aggressive or especially polite behavior), or to measure certain quantitative aspects, such as the number of inputs until a scenario has been completed or the time needed to send an input. The *collecting agent* collects the results of the single agents and merges them. The results of the analysis influence the answering behavior of the chatbot as well as the player's individual score, which is calculated by the *score agent*. If the collecting agent finds something that requires an immediate reaction of the chatbot (e.g. aggressive or rude behavior), it modifies the user messages by replacing it by a keyword that triggers an appropriate reaction of the chatbot. Based on the combination of the selected sentence opener, the free text input, and the analysis results, the chatbot (*AIML processor*) creates the customer's response to the user message. SQLSpaces enables the logging of the whole conversation and all important game data, which is needed to provide the replays and statistics embedded in the group reflection tool. The *serialization agent* is responsible for this task.

4 Evaluation

Our goal for the evaluation was to investigate if the developed scenarios and the chatbots qualify for a real training situation. In a mixed-method study, qualitative and quantitative methods have been used to gain insights regarding the playability and perception of the chatbots.

4.1 Experimental Design

Population. 20 participants (average 26.05, SD = 7.99, 15 females, 5 males) participated in the study. 80% of the participants had a university-entrance diploma, 5% a vocational diploma, and 15% already had a university degree. 4 out of 20 participants indicated that they have experience in customer complaint management. The participants were recruited from the University of Duisburg-Essen campus, as well as through announcements and social media.

Procedure and Data Collection. After the briefing, the participants were asked to fill in a descriptive questionnaire. They were introduced to the game and started playing the first scenario, which was design to allow the participants to familiarize with the user interface. After completing it, they played either the second or the third scenario. The distribution was randomized. Every participant received a checklist presenting some basic rules in regard to complaint handling based on the complaint management checklist from Stauss and Seidel [17]. The gaming session was followed by answering several post-experiment questionnaires to collect the experiences and perceptions during the gaming session. Afterwards, a short interview gave the participants the opportunity to describe their experience with the game in their own words before the debriefing commenced. The whole experiment took roughly one hour per participant.

4.2 Goals and Hypotheses

Our main goal for this study was to investigate whether the developed scenarios are perceived as realistic and can be used in real training situations. In addition, we wanted to examine if the developed chatbots behave as desired and whether their conversation style is influencing the users' perception of the chatbots. We assumed that people with prior knowledge in customer complaint management would perform better than people without prior knowledge. To answer these questions and examine the validity of our assumptions, three main hypotheses were established:

1. Participants who play the second scenario ("compromising") achieve different results in the *game experience questionnaire* (GEQ) dimensions *tension*, *negative affect*, and *challenge* than participants who play the third scenario ("dominating").
2. Participants who play the second scenario ("compromising") achieve different results in the Holtgraves questionnaire dimensions *comfortable*, *thoughtful*, *polite*, *responsive*, and *engaging* than participants who play the third scenario ("dominating").
3. Participants with prior experience/knowledge in complaint management achieve better results than participants without prior experience/knowledge.

4.3 Method of Analysis

Subjective Measures. The questionnaire was composed of four different parts. The first part included the collection of demographic data and prior experience in complaint management. For the assessment of game experience, the GEQ [18] was used in the second part. It covers the seven dimensions of game experience*: immersion, tension,*

competence, flow, positive affect, negative affect, and *challenge*. A total of 42 items have been rated on a five-point scale. In the third part, the playability and perceived usefulness of the scenarios were measured with the questionnaire developed in the context of the evaluation of the game ENACT (hereinafter referred to as the *EduTechRPG questionnaire).* Because CuCoMaG is a text-based role-playing game without avatars, only nine of the initial thirteen items were used. The items have been rated on a five-point scale. To measure the perception of the chatbot, a questionnaire concerning the human-like qualities of the chatbots was used (developed by Holtgraves, Ross, Weywadt and Lin [19]). The conversation with the chatbot was to be rated with seven bipolar adjective pairs. Each pair has been rated on a nine point scale with the positive adjective corresponding to the value nine and the negative to the value one. To gain qualitative data from the participants, the following questions were asked in a short qualitative interview: (1) How did you experience the conversation with the chatbot? (2) How did you act during the dialog? (3) How did you behave when problems occurred during the ongoing conversation?

Objective Measures. To complement the results from the questionnaires and the interview, the dialog scripts were evaluated in regard to the answer quality of the chatbot. The answers were assigned to one of three categories: *constructive*, *comprehensible,* and *nonsensical*. The categorization relied on human coding based on a clear operational classification inspired by Shawar and Atwell [20]. For each scenario, we determined the mean number for each category. In addition, the frequency of uses for each sentence opener was counted to estimate which were used frequently, rarely, or not at all. With regard to hypothesis 3, the success in the game is operationalized in two different ways. First, the relative score is determined by dividing the total score as calculated by the internal scoring system of the game by the number of text inputs. Rude, aggressive, or inappropriate behavior as well as long pauses reduce the score while polite behavior and fast response times increase it. The second indicator is the number of inputs because it is assumed that a rapid completion of the scenarios indicates an effective complaint management. The result from the two scenarios a participant completed has been added up to an overall value.

4.4 Results

GEQ. In order to assess the overall game experience, the arithmetic mean of the values from all participants are considered. As can be seen in Table 1, all dimensions with the exception of *negative affect* are above average, with the dimensions *flow*, *positive affect*, and *immersion* achieving the highest values. In order to provide a more differentiated view on the different scenarios, the mean values of the two data sets (second scenario and third scenario) are compared by applying a t-test. Levene's test only becomes significant for the dimension *tension* ($p = -0.37$), so we used the corrected values for this, whereas variance homogeneity can be assumed in the other cases. Overall, only the difference regarding the dimension *negative affect* is significant according to the t-test ($t(18) = -3.10$, $p = .006$).

Table 1. Results of the GEQ (0 to 4 scale)

	M (scenario 1 + 2)	M (scenario 1 + 3)	M (total)	SD	Min	Max
Immersion	3.18	2.97	3.08	0.74	1.50	4.17
Tension	1.85	2.45	2.15	0.88	1.00	3.83
Competence	3.22	2.62	2.92	0.92	1.00	4.50
Flow	3.35	3.08	3.22	0.64	1.83	4.50
Positive affect	3.43	3.10	3.27	0.81	1.17	4.33
Negative affect	1.45	2.35	1.90	0.78	1.00	3.67
Challenge	2.75	2.78	2.77	0.48	1.83	3.67

EduTechRPG Questionnaire. The arithmetic mean values for all items of the EduTechRPG questionnaire are above average, while the items "The scenarios deal about real-life situations" (M = 4.30 of 5), "The information given are useful and clear" (4.15 of 5), "The agents are behaving differently in each scenario" (4.10 of 5), and "I found it easier to negotiate with some agents than others" (4.00 of 5) achieve the highest values (Table 2). 85% of the participants stated that they would be willing to play the game again with different scenarios and characters. All participants who declared that they would not play the game again had played the third scenario ("dominating").

Table 2. Results of the EduTechRPG questionnaire (1 to 5 score)

	M	SD
The conversation with the agents is realistic	3.35	1.18
The user interface is intuitive and good-looking	3.50	0.83
The information given are useful and clear	4.15	0.67
The scenarios deal about real-life situations	4.30	0.66
The agents are behaving differently in each scenario	4.10	0.72
I found it easier to negotiate with some agents than others	4.00	0.65
I am motivated to negotiate even with the toughest agent	3.85	1.14
I find the overall experience with the CoCoMaG game positive	3.90	1.07
Would you play this game again with different scenarios and characters?	85% Yes 15% No	–

Perception of the Chatbots. The results regarding the perception of the chatbots of scenarios 1 and 2 show that they are perceived as especially *responsive* (M = 6.50), *comfortable* (M = 6.20 of 9), *polite* (M = 6.20 of 9), and *skilled* (M = 6.10 of 9). The mean values of the items *human* (M = 5.90 of 9), *thoughtful* (M = 5.90 of 9), and *engaging* (M = 5.70 of 9) are above average, too. In comparison, the chatbots of the scenarios 1 and 3 reach a value slightly above average on the item *skilled* (M = 5.10 of 9), whereas the values of the items *engaging* (M = 3.90 of 9), *responsive* (M = 3.80 of 9), *thoughtful* (M = 3.70 of 9), and polite are far below average. Again, a t-test was done in order to compare the mean values of the two groups. Since Levene's test does

not show any significant results, variance homogeneity can be assumed. The t-est shows significant differences for the items *thoughtful* ($t(18) = 3.75$, $p = .001$), *polite* ($t(18) = 3.95$, $p = .001$), *responsive* ($t(18) = 4.26$, $p < .001$), and *engaging* ($t(18) = 2.22$, $p = .039$).

Evaluation of Chat Protocols. Concerning the answer quality derived from the analysis of the chat protocols, the results show that the number of *constructive* bot responses is the highest for all scenarios. The number of *comprehensible* bot responses is lower than the number of *constructive* bot responses but higher than *nonsensical* bot responses (Table 3). The biggest number of the *comprehensible* (but not *constructive*) bot responses are default outputs, which were implemented for each sentence opener in case no input match could be found for the free text part of a chat message. These default outputs are supposed to show the players that the chatbot did not fully understand the message while still being aware of the context, and to encourage them to rephrase the message. The smallest part of the three answer categories in each scenario form the *nonsensical* answers. This category includes answers that either did not fit the player's input or were semantically correct but did not make sense in the context of the scenario.

Table 3. Mean values of the answer categories in the three different scenarios

	Scenario 1	Scenario 2	Scenario 3
M (constructive)	9.60 (SD = 4.39)	17.50 (SD = 4.74)	16.10 (SD = 6.26)
M (comprehensible)	3.65 (SD = 4.28)	5.40 (SD = 5.82)	5.70 (SD = 5.42)
M (nonsensical)	0.85 (SD = 0.99)	1.90 (SD = 2.03)	0.40 (SD = 0.52)

Differences in Success Depending on Prior Experience. For both indicators of the variable *success in game* (relative score and number of inputs) a t-test for independent samples was conducted. The two test groups are "participants with prior experience in complaint management" ($n = 4$) and "participants without prior experience in complaint management" ($n = 16$). The t-tests do not show significant results.

Frequency of Sentence Openers. Sentence openers that could be used to obtain information from the customers were the ones used most frequently. Those included "Tell me …" ($M = 3.50$) or "Please describe…" ($M = 3.90$). The sentence opener "I am sorry…" was the one used most frequently both in the third scenario ($M = 4.80$) as well as overall in the game ($M = 3.75$). Sentence openers that were not or barely used are "I cannot do that…" ($M = 0.00$) and "What do you think about…" ($M = 0.05$). Overall, the frequency of all sentence opener increases from the first to the second scenario that was played.

Qualitative Interviews. The results of the qualitative interviews on conversation perception vary greatly. Five participants stated that they attribute a *high degree of difficulty* to the scenarios. All of them played the third scenario and had no experience in complaint management. Three other participants found the scenarios *pleasant* and *uncomplicated*. Seven participants found the use of the sentence openers to be *inhibitory* and *restrictive*. All participants described their own behavior in the chat conversations as *polite*, while ten of them stated that they behaved in a *problem-oriented* or *solution-oriented* manner. The problem-solving approaches of the participants are distinguished

by four main approaches: Ten participants stated that they had rephrased their input in case a chatbot did not understand or did not provide meaningful answers. Eight participants tried to phrase their input with other sentence openers. Five participants reported having phrased completely new messages, and four that they were trying to repeat the same message. Some participants reported several of these solutions.

4.5 Discussion

Hypothesis 1 could be partially confirmed. There were significant differences in the GEQ dimension *negative affect* but not in the dimensions *tension* and *challenge* between the participants that completed the second or third scenario. The lack of significant results could be caused by methodical conditions like the small number of participants or the experimental design, as the participants were asked to evaluate the perception of both played scenarios combined.

Hypothesis 2 could be partially confirmed as well. Participants who played the second scenario showed significant differences in the dimensions *thoughtful*, *polite*, *responsive,* and *engaging*. This result supports the successful character design of the chatbots. As predicted, there were no significant differences in the dimensions of *human* and *skilled*, which suggests that there is only a difference in the chatbots' conversation style but not in the quality of their implementation. The dimension *comfortable* was not significant between the scenarios. This could also be caused by methodical conditions.

Hypothesis 3 could not be confirmed. We expected to find differences in the performance between participants with and without prior experience in complaint management, but statistical tests were not possible because of the small sample size ($n = 4$).

In general, the results of the GEQ and the EduTechRPG questionnaire indicate that the scenarios are perceived as realistic and that the game experience is quite positive. Especially good results have been achieved in the items "The scenarios deal about real-life situations", "The information given are useful and clear", "The agents are behaving differently in each scenario", and "I found it easier to negotiate with some agents than others" of the EduTechRPG questionnaire, which could be the result of the sophisticated design based on psychologically supported models underlying the developed scenarios [1, 8, 10, 11]. Only 15% of the participants stated that they would not want to play the game again with other characters and scenarios. All of these participants played the third scenario, which had significantly higher values in *negative affect*.

The participants' response behavior is consistent with the results of the chat feedback. All of the participants reported having been polite, which corresponds to the results of the analysis agents. Rephrasing was most commonly used as a solution to comprehension problems, which may have been supported by the tailored default responses of the bots.

The high frequencies of the sentence openers for the collection of information can be explained by their very variable possibilities of supplementation. The frequent use of the sentence opener "I am sorry" is not surprising, since apologies are phrases that are almost always suitable as a reaction and are very clearly associated with polite behavior [10]. Some of the participants claimed that they had problems expressing

themselves and creating sentences based on the predefined sentence openers. This feedback should be used to improve and expand the offered set of sentence openers in order to support the players and provide more and better opportunities to express themselves. One other important result of the evaluation is that the dialog scripts have potential for improvement and that they need to be expanded, e.g., by covering more synonyms and unexpected inputs.

A major limitation of this preliminary study is the sample size (especially for participants with experience in customer complaint management), which may be the reason, why hypotheses 1 and 2 could only be partially confirmed and hypothesis 3 could not be statistically tested. The experiment needs to be repeated with a considerably larger sample in order to allow for generalization.

5 Conclusion and Future Work

Although there are a number of educational games and simulations addressing the training of specific social skills, no approach exists explicitly targeting the training of customer complaint management strategies. In this paper, we have presented a novel and innovative approach towards providing a 2D role-playing environment for the training of customer complaint management in the form of an educational game that adequately fits and supports the training of complaint conversations. There is no existing approach explicitly targeting the training of customer complaint management strategies. Our system environment can be naturally extended with new customer cases representing challenges that focus on a specific subset of skills each and thus allows for organizing the learning process as a sequence of cases. The evaluation of the game showed on the one hand that the idea and approach of the game in general were assessed positively and most of the participants considered it worthwhile to play the game several times. On the other hand, the evaluation revealed problems, especially with the application of the predefined sentence openers, which will be adapted to further improve the game flow.

Although the hypotheses have been only partially confirmed, it could be validated that the discussion style of the chatbots is influencing the players' perception of the dialog partner and the game experience, which underlines the successful design of the chatbots. Due to the small number of participants, a generalization cannot be made, but the results are promising and should be expanded in larger studies after a revision and extension of the prototype. More scenarios could be added to increase the variety and to offer more levels of difficulty. To evaluate the chatbot in its intended field of application, it would be reasonable to test the training scenarios directly in companies that might use this kind of training software for the professional training of their employees.

References

1. Cho, Y., Im, I., Hiltz, R., Fjermestad, J.: An analysis of online customer complaints: implications for web complaint management. In: Proceedings of the 35th Annual Hawaii International Conference on System Sciences, pp. 2308–2317. IEEE (2002)

2. Liljander, V., Strandvik, T.: Emotions in service satsifaction. Int. J. Serv. Ind. Manag. **8**(2), 148–169 (1997)
3. Heung, V.C.S., Lam, T.: Customer complaint behaviour towards hotel restaurant services. Int. J. Contem. Hospitality Manag. **15**(5), 283–289 (2003)
4. Dell'Aquila, E., Marocco, D., Ponticorvo, M., Di Ferdinando, A., Schembri, M., Miglino, O.: Educational Games for Soft-Skills Training in Digital Environments. Springer, Switzerland (2017)
5. Knode, S., Knode, J.-D.: Using a simulation program to teach leadership. In: Proceedings of the 2011 ASCUE Summer Conference, pp. 86–92 (2011)
6. Emmerich, K., Neuwald, K., Othlinghaus, J., Ziebarth, S., Hoppe, H.U.: Training conflict management in a collaborative virtual environment. In: Herskovic, V., Hoppe, H.Ulrich, Jansen, M., Ziegler, J. (eds.) CRIWG 2012. LNCS, vol. 7493, pp. 17–32. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33284-5_2
7. Ziebarth, S., Kizina, A., Hoppe, H.U., Dini, L.: A serious game for training patient-centered medical interviews. In: 14th International Conference on Advanced Learning Technologies, pp. 213–217. IEEE (2014)
8. Fornell, C., Wernerfelt, B.: A model for customer complaint management. Mark. Sci. **7**(3), 287–298 (1988)
9. Stauss, B., Seidel, W.: Beschwerdemanagement: Unzufriedene Kunden als profitable Zielgruppe. Carl Hanser Verlag, Munich (2014)
10. Chase, R.B., Dasu, S.: Want to perfect your company's service? use behavioral science. Harvard Bus. Rev. **79**, 78–84 (2001)
11. Stauss, B.: The dimensions of complaint satisfaction: process and outcome complaint satisfaction versus cold fact and warm act complaint satisfaction. Managing Serv. Qual. Int. J. **12**(3), 173–183 (2002)
12. Rahim, A., Bonoma, T.V.: Managing organizational conflict: a model for diagnosis and intervention. Psychol. Rep. **44**(3), 1323–1344 (1979)
13. Jonassen, D., Mayes, T., McAleese, R.: A manifesto for a constructivist approach to uses of technology in higher education. In: Duffy, T.M., Lowyck, J., Jonassen, D.H., Welsh, T.M. (eds.) Designing Environments for Constructive Learning, pp. 231–247. Springer, Heidelberg (1993). https://doi.org/10.1007/978-3-642-78069-1_12
14. Wallace, R.: The elements if AIML Style. ALICE AI Foundation (2004)
15. Ghose, S., Barua, J.J.: Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor. In: 2013 International Conference on Informatics, Electronics & Vision (ICIEV), pp. 1–5. IEEE (2013)
16. Weinbrenner, S., Giemza, A., Hoppe, H.U.: Engineering heterogenous distributed learning environments using TupleSpaces as an architectural platform. In: Proceedings of the 7th IEEE International Conference on Advanced Learning Technologies (ICALT 2007), pp. 434–436. IEEE (2007)
17. Stauss B., Seidel, W.: Complaint management. In: Introduction to Service Engineering, pp. 414–432. Wiley, New Jersey (2010)
18. Ijsselsteijn, W.A., De Kort, Y.A.W., Poels, K.: The game experience questionnaire. Technische Universität Eindhoven (2013)
19. Holtgraves, T.M., Ross, S.J., Weywadt, C.R., Lin, H.T.: Perceiving artificial social agents. Comput. Hum. Behav. **23**(5), 2163–2174 (2007)
20. Shawar, B.A., Atwell, E.: Chatbots: are they really useful? LDV Forum **22**(1), 29–49 (2007)

6 CONCLUSION

This chapter presents additional research results of a study on the customer complaint management scenario conducted in 2020, which have not been published yet. Furthermore, the main findings of this thesis will be summarized and future perspectives elaborated.

6.1 ADDITIONAL RESULTS

Following the experiments described in Chapters 3 and 4, an additional study on the customer complaint management scenario (CuCoMaG) has been conducted, this time with emphasis on the post role play group reflection phase. This subsection summarizes the (so far unpublished) results of the study.

In 2020, the whole training system was re-designed, amended and polished in regard to the following aspects:

- The design was improved to foster easy and intuitive use (Figure 3).

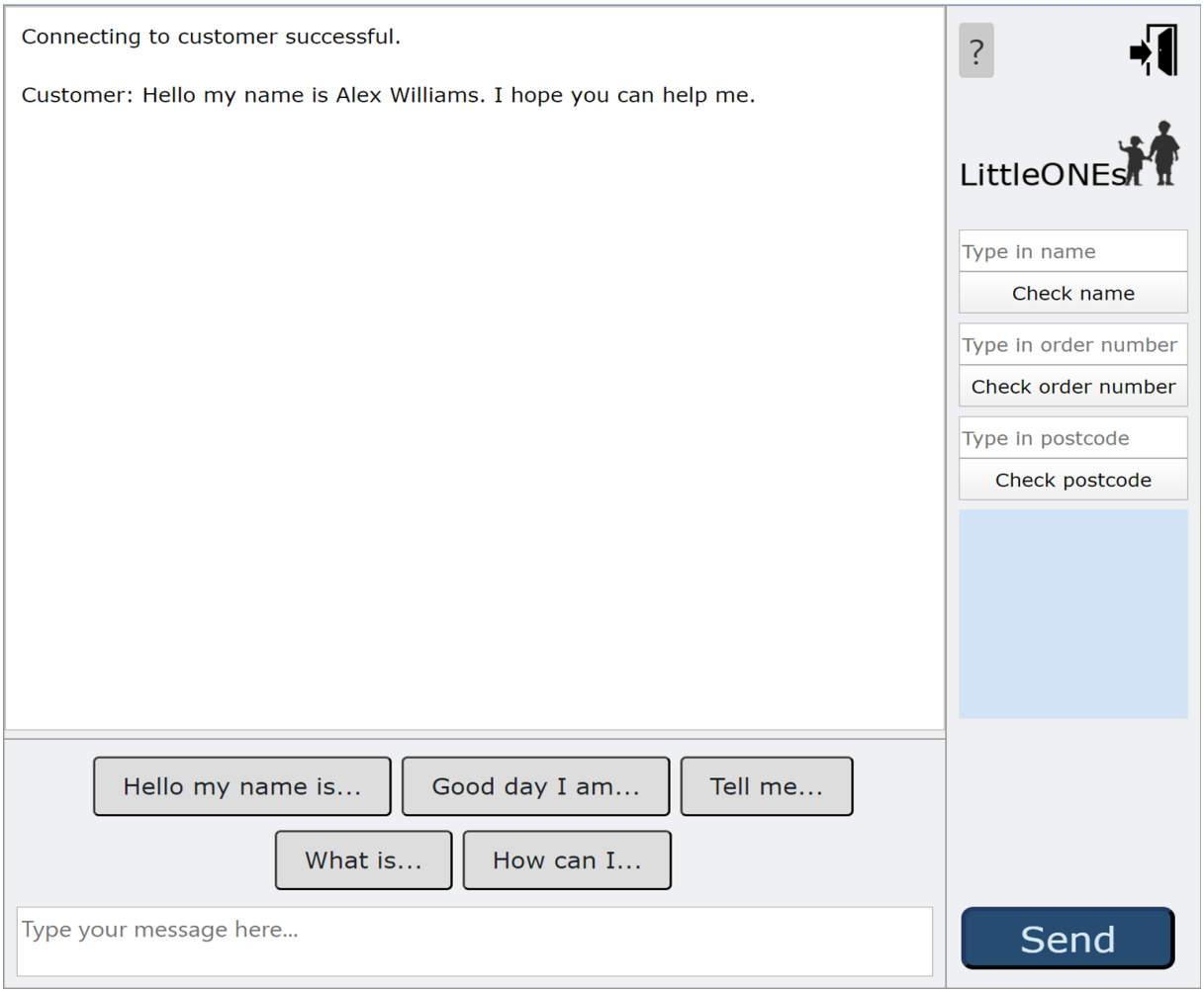

Figure 3: Amended chat interface

- The rather long textual introduction in the beginning of the game was transformed into a cartoon-like intro (Figure 4).

Figure 4: Cartoon-like introduction

- An interactive tutorial was added for introducing the user to the interface and its capabilities before being engaged in the actual scenarios (Figure 5).

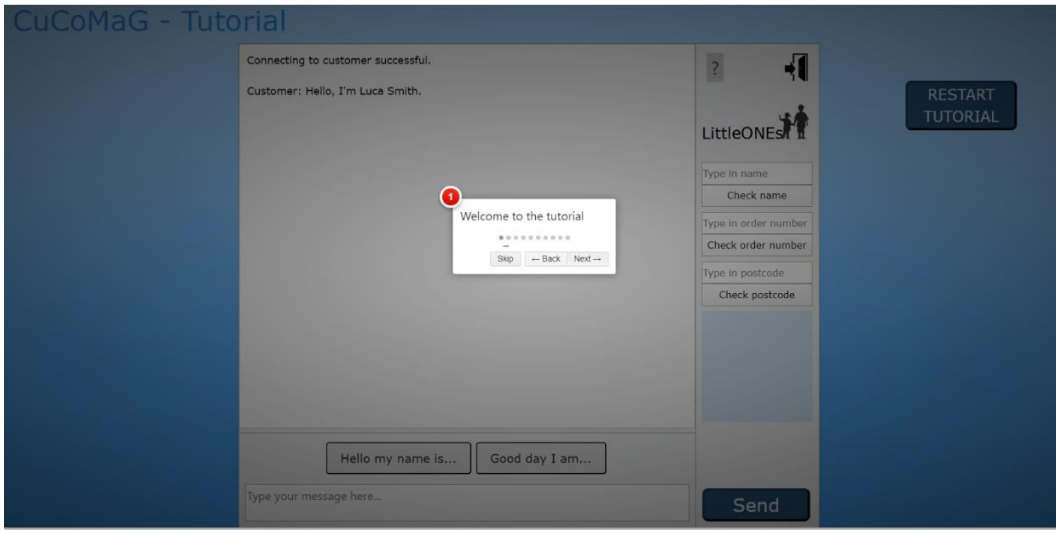

Figure 5: Interactive tutorial

- The existing scenarios were improved upon and polished, and one additional scenario was added.
- A virtual tutor was added to accompany and guide the user through the course of the game and provide support when needed (Figure 6).

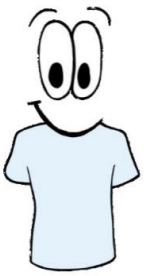

Figure 6: The virtual tutor in CuCoMaG

- A questionnaire for self-assessment was integrated to allow for a comparison between the perception of the own performance and the actual adduced performance (Figure 7).

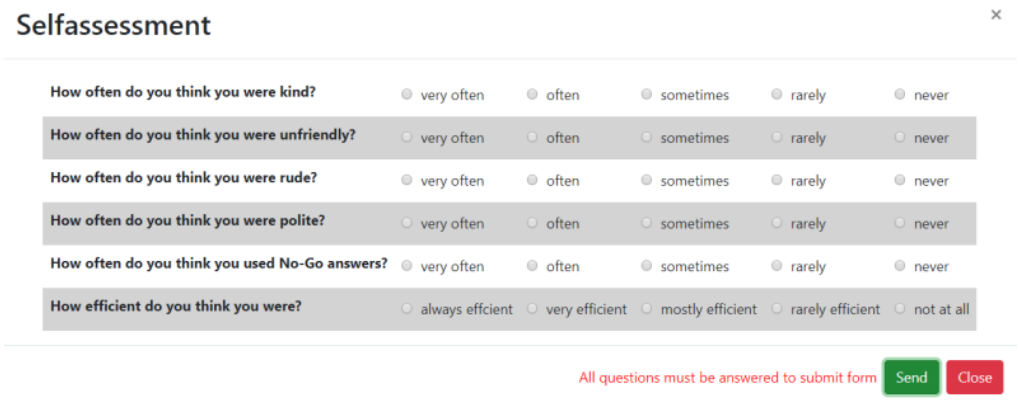

Figure 7: Embedded Self-assessment questionnaire

- The group reflection tool was completely re-designed, improved and amended (Figure 8).

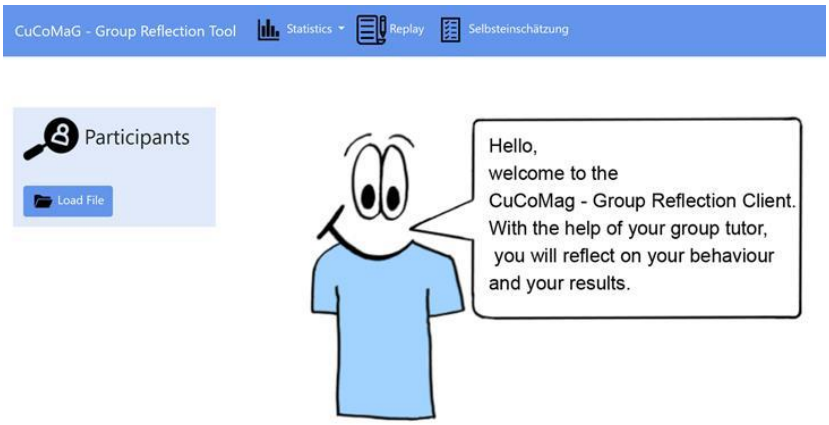

Figure 8: New group reflection tool

The main goals of the study were to find out whether the new design is appealing to the users, whether users who already had prior experience with customer complaints perceive the game to be immersive, and especially whether the new group reflection tool can support reflection processes. This is the first study, which primarily focuses on the (post role play) reflection phase.

In order to explain the evaluation procedure and results, it is necessary to first describe the functionalities of the re-worked group reflection tool. As shown in Figure 8, the new user interface of the group reflection tool includes a navigation bar at the top and a list of participants at the left. This list shows the current participants of the group reflection session and their total scores achieved in the previous gaming session. With help of the button at the bottom, it is possible to add data of as many users as desired. The navigation bar contains the three main functionalities “statistics”, “replay”, and “self-assessment”. The first menu point *statistics* offers different visualizations of the users’ performances in form of bar or line charts. The bar chart (Figure 9) shows how many times the selected characteristics were used by each

of the players. Any characteristic combinations are possible, and the chart dynamically adapts to selections and deselections. The players can be selected (or deselected) in the chart itself (at the bottom).

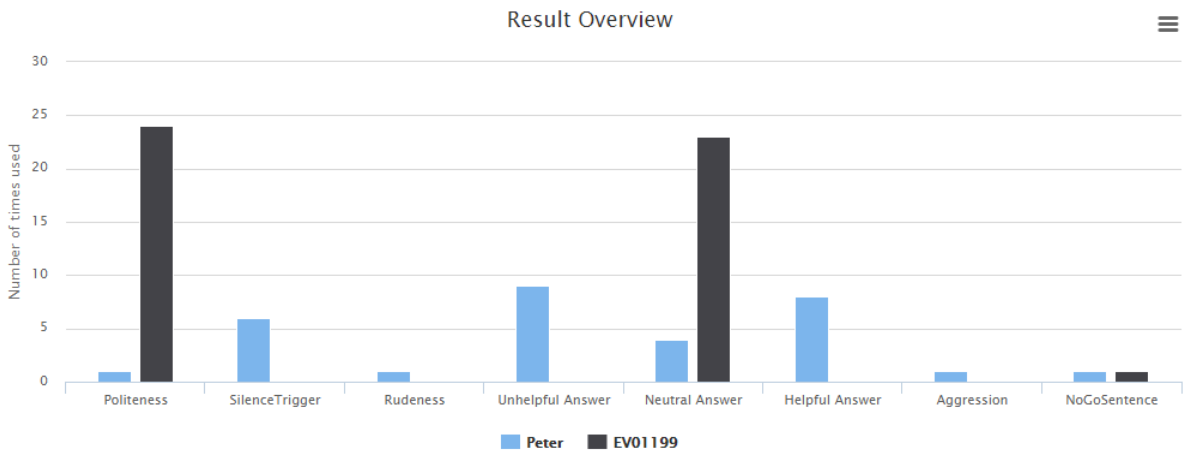

Figure 9: Bar chart

The line chart works in a similar way. It displays the development of the players' score in the course of the whole chat conversation. Upon selecting one or more characteristics (e.g., rudeness), dots in different shapes appear on the line chart, representing the messages where those specific characteristics applied. By hovering the mouse over a message, a small box (tooltip) appears, which shows more detailed information, including the score and the characteristics which applied in that specific message. The contents of the tooltips, including the texts and the corresponding colors, are generated dynamically based on the characteristics selected by the user. Trainers can directly switch from the line chart to a certain point of interest in the chat history by clicking on a point in the chart. The display changes to the conversation view and highlights the selected message.

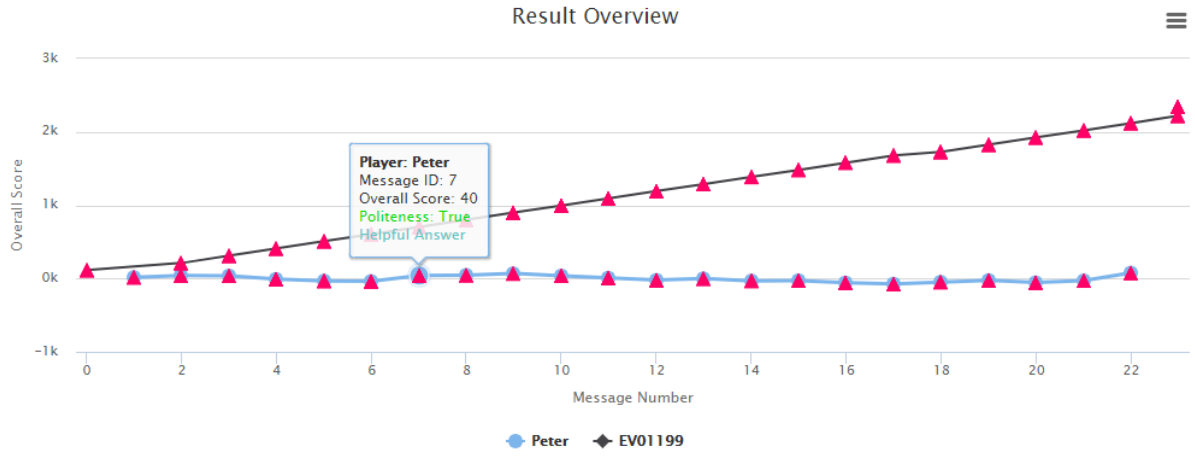

Figure 10: Line chart

The second menu point "replay" allows the user to open chat transcripts of the players' gaming sessions, which makes it possible to go through the course of the complaint conversations again retrospectively. In the selected replay, the respective player's inputs are augmented with helpful system-generated annotations based on the results of the analysis

agents. The replay was extended by additional functionalities: Similar to the charts, it is now possible to view more than one dialog in parallel, i.e., the user can view up to two replays simultaneously. The two replays are displayed directly next to each other and easy navigation and systematic comparison between them is possible. In this way, the players can quickly learn how to uncover their own mistakes by seeing how another participant of the group reflection session handled the customer complaint process. Another function that has was added within the scope of the re-design is the keyword search. With help of this function, it is possible to search for specific words within the chat protocols for easy navigation.

The third option, “self-assessment”, allows a contrasting juxtaposition of the results of the self-assessment survey each player completed right after the gaming session and the actual result in form of a grouped bar chart showing the different parameters like rudeness, unfriendliness, politeness, etc. as well as the respective frequencies of occurrence within the course of the gaming session. The absolute frequencies are mapped to the discrete values “rarely”, “sometimes”, “often”, and “very often”, which have been used in the game’s self-assessment form. This form of comparison is supposed to support the reflection process. According to Dewey and Schön, the beginning of a reflection process is usually triggered by a discrepancy experience (Dewey, 1933) (Schön, 1983) and this is what the self-assessment function is drawing on. If persons perceive a difference between their own assessment and their actual behavior, they begin to locate the difficulties that led to the discrepancy (Dewey, 1933). This process can be supported by, for example, checking the replay, which helps to identify specific behaviors and actions responsible for the results. This way, the tool can help to find approaches towards a solution.

The following passages summarize the results of the study conducted with the re-worked prototype.

Research questions and hypotheses. The main goals of the study were to test whether the design and gameplay improvements are well received by the users and to evaluate whether the group reflection environment is a suitable tool for supporting reflection processes and if users can work with the provided functionalities. Concretely, one aim was to find out whether the participants (i.e., more than half of them) find the new design better than the old design (hypothesis 1). It was also tested whether participants who had already come into contact with customer complaints find the game immersive (hypothesis 2). In addition, it was examined whether the participants (i.e., more than half of them) understood and used the self-reflection features in the group reflection tool (hypothesis 3).

Experimental Design. 33 participants (average 26.88, 19 females, 14 males) participated in the evaluation study, from which 22 had already contacted a customer complaint center in the past and 4 had already worked in customer complaint management. After the introductory comic, all participants completed the interactive tutorial and played the second scenario. Like in the previous study, all participants received a checklist including basic rules for handling complaints (Stauss & Seidel, 2010). After the gaming session, the participants filled out several post experiment questionnaires. These included the GEQ to retrieve their experiences and perceptions during the game as the first questionnaire. After that, a printout with the new and old design was presented to the participants, who then had to fill out a self-designed

questionnaire about the design. On a five-point Likert scale, the participants had to rate specific design aspects (e.g., “I like the layout better in the new design”, “I like the old logo better.”, “The level display in the new design has helped me know where I am in the game.”). After that, there was a switch to the group reflection tool, in which the participants had to complete four tasks. These four tasks were:

1. *Bar Charts*: The participants were asked to compare the results of a given user with their own results and to find differences in the performance. In addition, they had to explain what these differences show them and what they could tell from the differences.
2. *Line Charts*: Here again, the participants were asked to compare the results of a given user with their own results and to find differences in the performance. The task is similar to the first one, only based on a different chart.
3. *Replay and search function*: First, the participants were asked to find utterances with specific annotations (#aggression and #rudeness) in two given chat protocols (one of them is the protocol of the current user and the other one a prepared sample. They were asked to copy the concerned message into the notepad and explain what these annotations mean and why specific messages are marked with these. Second, the participants were asked to use the search functions to find differences or abnormalities between the two given replays and comment on why the concerned messages are noticeable and in what way they differ from each other. Third, the participants were asked to convert their written notes into a PDF file.

The whole process – including the answers to the given tasks – was observed and documented by the experimenter. The perceived usability of the group reflection tool was assessed with a short version of the *Usability Experience Questionnaire* (Schrepp, Hinderks, & Thomaschewski, 2017). Finally, the participants answered some demographic questions and whether they had ever been in contact with customer complaints (and if so, how often), and whether they had worked in the field of customer complaint management.

Method of analysis. For the validation of the first hypothesis, which states that more than half of the subjects prefer the new design to the old design, the cumulative values of the individual subjects for the design questionnaire were determined. Then, the mean of the sum scores of all subjects was calculated. For the confirmation of the hypothesis, with a maximum sum value of 75, at least a mean value of 37.5 was necessary. For the second hypothesis, which states that subjects who have already come into contact with customer complaints find the game immersive, the participants who had indicated in the questionnaire that they had already contacted a customer complaint center or had already worked in customer complaint management were first filtered out. The GEQ was then evaluated from these 22 participants, with a special focus on immersion. For this purpose, the mean values of all subjects were evaluated for all questions with a focus on immersion. Finally, the third hypothesis was tested, which states that the functions for reflection in the group reflection tool were understood and used by more than half of the subjects. For this purpose, the data of the respective experimenters were evaluated. For each function, it was checked whether it was used and understood in a meaningful way. Thus, for each function, a maximum value of 33 subjects could be achieved who understood and used the function.

Evaluation results. Regarding the first hypothesis, the evaluation showed a mean value of 59.45, which is above the critical value. The range of 41 - 69 is also still above the critical value. Thus, the first hypothesis was confirmed. Although not all of the design decisions made were equally well received by the participants, it can nevertheless be stated that none of the changes were predominantly evaluated negatively. Regarding the second hypothesis, the statistics show that although only one question is below the minimum mean of 2.5 to be achieved ("it felt like a rich experience" = 2.45), the remaining questions are also only slightly above the mean to be achieved. Thus, the hypothesis is not rejected here either, but it should be viewed very critically. On the one hand, the subjects felt mostly ingenious and interested in the story of the game, but on the other hand, the game did not feel like a rich experience for the majority of the participants. One possible reason for this could be the small sample size. In addition, some participants were unable to advance to the third phase of the level without a hint from the experimenter, so the experimenter had to intervene. Furthermore, not all player utterances were understood by the chat bot, which could be due to the bot's lack of comprehension on the one hand, and grammatical and orthographic errors by the participants on the other. As a result, the customer conversation may have been perceived as unrealistic by the players. These aspects may have compromised the immersion. Regarding the third hypothesis, the results show that the functionality of the line and bar charts as well as the replay were understood and used by all 33 subjects. In the case of the search function of the replay, 29 of 33 subjects used the search function in an appropriate way. This clearly confirms the third hypothesis. Overall, all hypotheses could be confirmed, although the confirmation of the second hypothesis should be viewed rather critically. In addition, although on the one hand all the hypotheses that were formulated were supported, the way they were formulated made it impossible to evaluate them by common statistical means, such as a t-test, which is why procedures such as summation were used. As a result, no statistically significant results could be found. However, if the procedures are applied correctly, this aspect should not make the results less meaningful.

In summary, this additional study focused on one aspect that has not been evaluated within the scope of the main publications included in this thesis: reflection support. While the other studies focused mainly on the role-playing part of the game environments, this study aimed to find out if the group reflection environment is a suitable tool for supporting reflection processes and users can work with the provided functionalities, and the results confirm this.

6.2 SUMMARY

This subsection summarizes the most important developments and findings of this work from different perspectives (design perspective, technical perspective, and evaluation perspective):

Serious games have been established as an efficient medium in education and professional training in recent years. They have the capability to be effective tools to promote learning and encourage behavioral change, and they constitute a vital instrument for a variety of education and training scenarios. The combination of the serious gaming approach with role-playing is particularly promising, as authentic simulated environments provide mobile, safe, and continuable settings for learners, in which they can assume roles in particular contexts,

explore new situations, and learn how to act and react without fearing consequences in the real world.

A special challenge for this kind of game is shaping the pedagogical outcomes, as the effects generally depend on post-role-play reflection. Without feedback and reflection, the transfer to real world situations cannot be ensured. Computer-supported analyses can help to track and evaluate the learners' performances, generate feedback, and provide structured recordings enriched with helpful features like integrated indexing, navigation instruments, search functions, and cross references between different media and data sources.

This thesis presented a technical and conceptual framework for the development of serious role-playing games designed for the training of specific social skills involving chatbots in dialog-centric settings. From the design perspective, the framework is characterized by three distinctive conceptual features:

1. Chat-like interaction with AI-controlled chatbots
2. Strict separation of the immersion phase (role-playing) and the reflection phase (either individually or in a group)
3. Adaptive feedback and individual performance analyses

From the technical perspective, the implementation is based on three main components:

1. Sophisticated dialog-modeling based on AIML and additional tricks resulting in chatbots that adapt to the player's behavior
2. A multi-agent blackboard system as the technical basis, which enables performance optimization due to parallel processing and allows to keep components independent from each other
3. Intelligent support for automated performance evaluation and feedback generation

The presented framework facilitates the development of serious virtual role-playing games in several ways. First, it provides not only a conceptual basis for the development, but rather a complete game structure including all relevant components. The framework provides all basic mechanisms needed to develop a chatbot-based serious role-playing game, such as the inter-operability between GUI, chatbots and agents through a tuple space and thus represents a kind of modifiable prototype. Second, it allows for tailoring and adapting a given basic architecture with very limited effort, which means that it can easily be adjusted to many different settings and use cases. In order to develop a new application scenario, the following components would need to be amended or added:

1. A specific GUI with sentence openers
2. A new set of AIML scripts (chatbots)
3. A possible modification and/or extension of the analysis agents

Compared to role-playing with real actors, the player has to invest much less in this kind of training and is not tied to the willingness of other people to play. The player can learn from mistakes without being observed by others or having to fear consequences in the real world, which is especially helpful for shy learners. However, the conversation in such environments is less natural because the player is interacting with a bot and is bound to a fixed number of

sentence openings. Furthermore, the chatbots' nonverbal feedback is limited by the technical possibilities.

In addition to the concrete framework, a set of general dimensions and challenges in the design of serious role-playing games for the training of social skills has been formulated:

1. Learning context (theoretical foundation and desired learning objectives)
2. Technical architecture and set-up (technologies and tools used for the technical implementation, basic system architecture)
3. Dialog models and degrees of freedom (communication structure with virtual characters ranging from predefined answers to sentence openers to free text input)
4. Feedback and scaffolding elements and mechanisms (essential for enabling reflection processes and offering support and guidance to the learners)
5. The relation between immersion and reflection (separation or integration of phases of immersion and reflection)
6. Collaboration support (number of human players involved in the game and possibility to enable collaborative learning)

This set of dimensions and challenges constitutes important aspects for the conceptualization, description, and comparison of such games and can serve as a solid base for game developers and designers in this field. It is very important to deal with these aspects during the design process of serious role-playing games and to make informed decisions after contrasting different approaches and strategies and selecting those that are most suitable in the given context.

Based on the presented framework, two different innovative case studies have been presented in this thesis, one scenario for the training of workplace-oriented conflict management (ColCoMa) and one for customer complaint management (CuCoMaG). Both scenarios were enhanced in several development cycles and evaluated regarding different aspects. The evaluation studies show that the developed scenarios are perceived as useful and realistic and may qualify for real training situations. Furthermore, all studies presented within this thesis underline the benefits and importance of mixed-method studies combining quantitative and qualitative data and, first and foremost, subjective and objective data. This approach allows a more complete and synergetic utilization and evaluation of data and enables the alignment of personal feelings, assessment, and motivation with objective criteria such as player performance.

The following compilation (Table 3) provides an overview of the different methods and instruments used in the presented studies:

	Quantitative Methods	Qualitative Methods
Subjective Methods	• *SUS Usability Test* (Brooke, 1996) • *Usability Experience Questionnaire* (Schrepp, Hinderks, & Thomaschewski, 2017) • *Game Experience Questionnaire* (Ijsselsteijn, De Kort, & Poels, 2013) • *Holtgraves Questionnaire* (Holtgraves, Ross, Weywadt, & Lin, 2007) • *EduTechRPG Questionnaire* (Dell'Aquila, et al., 2017) • *Interpersonal Attraction Questionnaire* (McCroskey & McCain, 1974) • *Group Awareness Questionnaire* (Mock, 2017)	• Pre-structured Interviews • Self-developed Questionnaires
Objective Methods	• Player Performance / success in game (system-generated performance analysis and feedback, total and relative score, completion time, achieved milestones) • Eye-tracking	• Observation • Evaluation of chat protocols

Table 3: Overview of mixed methods and instruments

The conducted studies show that specific methods can be combined particularly well as they complement each other effectively. Eye-tracking, for example, turned out to be particularly suitable as an accompanying method for analyzing collaborative aspects in the context of serious role-playing games, and can be combined well with qualitative objective measures (like the human evaluation of chat protocols based on clear operational classification schemes) as well as subjective measures in order to compare and contrast the subjective assessment of players with objective data. In general, it can be concluded that the combination of at least one subjective and one objective method is advisable, as well as the combination of quantitative and qualitative methods.

It has also been shown that player performance is an important objective measure that is required for several purposes (e.g., to compare the performance of players with and without prior experience in the area of interest or to relate it to certain experimental conditions). Player performance can be operationalized based on different parameters (e.g., system-generated performance analysis and feedback, total and relative score, completion time, achieved milestones). Ideally, the operationalization should involve both data tracked and/or generated by the system and human evaluation based on clear operational classification schemes.

6.3 FUTURE PERSPECTIVES

This final subsection elaborates on future perspectives for the research field based on this thesis.

6.3.1 Further developments

In general, and based on the proposed framework, more scenarios and use cases could be developed in various fields of application. The framework has the potential to be applied to numerous areas of application of professional training in the field of social skills. The flexible architecture and the clear conceptual set-up allow for the system being easily tailored to other scenarios without major implementational effort. Thus, it may serve as a scaffold and enable other researchers to build upon this work and develop a range of serious role-playing games tailored to the needs of the respective application scenario and target group.

The promising results obtained in the presented studies could be expanded in larger studies and applied to various contexts. To evaluate serious role-playing games in their intended fields of application, it would make sense to test the training scenarios directly in companies that might use this kind of training software for the professional training of their employees. It would be particularly sensible to observe and analyze group reflection situations, because this was not part of the existing evaluation studies yet.

In regard to the aspect of collaboration, further studies are needed to identify more factors that are influencing collaboration quality to improve collaborative processes in serious role-playing games. In Chapter 3, the convergence of visual foci of attention (measured through eye-tracking) has been identified as one of these factors, but several other factors might have an influence as well. This aspect is especially interesting for future research in this field as there are still many open issues – in collaborative role-playing games specifically and learning environments in general – to improve learning and collaboration.

Of course, the framework can be further enhanced. As of now, scaffolding, adaptation, and personalization are incorporated in the framework only to a limited extend. Augmenting these dimensions therefore proposes a significant challenge for future research. Future work could also include chat awareness tools, which, for example, indicate if a chat message has been read or if the dialog partner is currently composing a message. Learner profiles would allow to track the progress of the learners and propose tasks with a suitable level of difficulty for them. Additional support mechanisms could be integrated for less experienced learners (with the option to deactivate them for more advanced learners).

6.3.2 New implementational approaches

One of the main challenges during the implementation was the realization of the chatbots' behavior, which leads to the general question whether AIML is a sufficiently rich language for the realization of intelligent chatbots like the ones used in the presented scenarios. Although AIML is certainly easy to use, its capabilities are fairly limited. AIML chatbots are passive in their nature and work purely based on pattern matching algorithms, which can be only overcome by using external triggers. Furthermore, AIML is not designed to support multi-user

conversations. That means, AIML chatbots cannot differentiate between two or more players or identify the sender of a message – in the presented framework, tricks were employed to work around this issue by, for example, introducing different sentence openers for different users. Modern chatbot technologies rely on advanced techniques, such as information extraction and machine learning, but these chatbots are difficult to implement and need to be trained on large datasets of question-answer pairs, which is not always possible. However, it appears reasonable to explore and test other chatbot technologies.

Another main implementational issue that arose in the context of this work is the use of sentence openers for the communication with chatbots. On the one hand, these pre-defined sentence fragments represent a scaffolding mechanism for the players and also simplify the interpretation of the player input enormously. Through the sentence openers, the conversation can be structured without strictly defining dialog sequences like in graph-based conversations. In addition, they allow defining of the general gist of a message (e.g., affirmation, rejection, asking a question), which limits the number of possible inputs and facilitates the interpretation of a message. On the other hand, the use of sentence openers limits the possibilities of expression and may push players into a certain direction. All in all, sentence openers seem to be a good compromise, which limit the possible inputs on the one hand (making it easier to understand the general gist of a message), but still offers players to formulate their own inputs and express themselves more freely.

6.3.3 Theory- and evidence-based serious game design

As described in Chapter 1.1.1, there is a broad interest in the potential of serious games as vehicles for learning and training, and the training of social skills is a specific domain of note. However, it can be observed that there is a general lack of evidence-based and theory-driven models for creating and evaluating serious games. But without a solid theoretical foundation and the use of well-proven learning strategies, design principles, and development methods,

a) the transfer of acquired knowledge and skills to the real world cannot be ensured and
b) it is impossible to assess the efficacy of serious games.

While several models and frameworks exist for the general design of serious games, most of these are only describing fundamental components and formal design approaches without concrete design or implementation guidelines. Furthermore, the are no models or frameworks specifically addressing the design of serious role-playing games. The framework developed within the scope of this thesis is trying to fill this gap by providing concrete conceptual and technical guidelines for the design and implementation of such games in order to foster a systematic approach and optimized efficacy and evaluability. The provided technical framework presents a solid basis for further development, which can be enhanced and optimized in order to make it even more adaptative and customizable.

In conclusion, the outcome of this thesis constitutes a step towards a systemized, evidence-based, and theory-driven design and implementation of serious role-playing games for the training of social skills in dialog-centric settings with virtual characters. Furthermore, the repository of mixed methods used in the conducted studies can be utilized for research in various areas in the landscape of serious games and learning environments evaluation.

7 REFERENCES

Abdul-Kader, S. A., & Woods, J. (2015). Survey on Chatbot Design Techniques in Speech Conversation Systems. *International Journal of Advanced Computer Science and Applications, 6(7)*, pp. 72-80. doi:10.14569/IJACSA.2015.060712

Annetta, L. A. (2010). The "I's" have it: A framework for serious educational game design. *Review in General Psychology, 14(2)*, pp. 105-113. doi:10.1037/a0018985

Augello, A., Gentile, M., & Dignum, F. (2016). Social agents for learning in virtual environments. *International Conference of Games and Learning Alliance* (pp. 133-143). Cham: Springer. doi:10.1007/978-3-319-50182-6_12

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* New Jersey: Prentice Hall.

Beale, I. L. (2011). *Video Games for Health.* New York: Nova Science Publishers.

Bellotti, F., Berta, R., De Gloria, A., & Primavera, L. (2010). Supporting authors in the development of task-based learning in serious virtual worlds. *British Journal of Education Technology, 41(1)*, pp. 86-107. doi:10.1111/j.1467-8535.2009.01039.x

Black, R. S., & Langone, J. (1997). Social awareness and transition to employment for adolescents with mental retardation. *Remedial and Special Education, 18*(5), pp. 214-222. doi:10.1177/074193259701800403

Boud, D., Keogh, R., & Walker, D. (1985). Promoting Reflection in Learning: A Model. In D. Boud, R. Keogh, & D. Walker (Eds.), *Reflection: Turning Experience into Learning* (pp. 18-40). London: Kogan Page.

Brackett, M. A., Rivers, S. E., & Salovey, P. (2011). Emotional intelligence: Implications for personal, social, academic, and workplace success. *Social and Personality Psychology Compass, 5*(1), pp. 88-103. doi:10.1111/j.1751-9004.2010.00334.x

Bradeško, L., & Mladenić, D. (2012). A survey of chatbot systems through a loebner prize competition. *Proceedings of Slovenian Language Technologies Society Eighth Conference of Language Technologies*, (pp. 34-37).

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. *International Conference on Internet Science* (pp. 377-392). Cham: Springer. doi:10.1007/978-3-319-70284-1_30

Braun, D., & Matthes, F. (2019). Towards a Framework for Classifying Chatbots. *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019). 1*, pp. 484-489. SCITEPRESS - Science and Technology Publications.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry, 189*, pp. 4-7.

Brusk, J., & Björk, S. (2009). Gameplay Design Patterns for Game Dialogues. *Proceedings of the 2009 Digital Games Research Association Conference.* London.

Charlop-Christy, M. H., & Kelso, S. E. (2003). Teaching Children With Autism Conversational Speech Using a Cue Card / Written Script Program. *Children, 26*(2), pp. 108-127.

Cheng, M.-T., She, H.-C., & Annetta, L. A. (2015). Game immersion experience: its hierachical structure and impact on game-based science learning. *Journal of Computer Assisted Learning, 31*(3), pp. 232-253. doi:10.1111/jcal.12066

Cheong, Y.-G., Khaled, R., Grappiolo, C., Campos, J., Martinho, C., Ingram, G., . . . Yannakakis, G. (2011). A computational approach towards conflict resolution for serious games. *Proceedings of the 6th International Conference on Foundations of Digital Games*, (pp. 15-22). doi:10.1145/2159365.2159368

Cho, Y., Im, I., Hiltz, R., & Fjermestad, J. (2002). An Analysis of Online Customer Complaints: Implications for Web Complaint Management. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* (pp. 2308-2317). IEEE. doi:10.1109/HICSS.2002.994162

Cho, Y., Im, I., Hiltz, R., & Fjermestad, J. (2002). An Analysis of Online Customer Complaints: Implications for Web Complaint Management. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* (pp. 2308-2317). IEEE.

Clayton, A. S. (2017). *Multiplayer educational role playing games (MPERPGs) and the application of leadership.* Ann Arbor: ProQuest Dissertations Publishing.

Coleridge, S. T. (1984). *Biographia Literaria.* Princeton: Princeton University Press.

Combs, M. L., & Slaby, D. A. (1977). Social-skills training with children. *Adv. Clin. Child Psychol., 1*, pp. 161–201. doi:10.1007/978-1-4613-9799-1_5

Corbett, A. T., Kenneth, R., Koedinger, J. R., & Anderson, J. R. (1997). Intelligent tutoring systems. In M. G. Helander, T. K. Landauer, & P. V. Prabhu, *Handbook of Human-Computer Interaction (Second Edtion)* (pp. 849-874). doi:10.1016/B978-044481862-1.50103-5

Corti, K. (2006). Games-based Learning; a serious business application. *Informe de PixelLearning, 34(6)*, pp. 1-20.

Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience.* New York: Harper and Row.

De Troyer, O. (2017). Towards effective serious games. *9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, (pp. 284-289). doi:10.1109/VS-GAMES.2017.8056615

Dell'Aquila, E., Marocco, D., Ponticorvo, M., Di Fernando, A., Schembri, M., & Miglino, O. (2017). *Educational Games for Soft-Skills Training in Digital Environments.* Cham: Springer. doi:10.1007/978-3-319-06311-9

Dewey, J. (1933). *How we think: a restatement of the relation of reflective thinking to the educative process.* Chicago: Henry Regnery Co.

Dewey, J. (1938). *Experience and Education.* New York: Collier Books.

DiCicco-Bloom, B., & Crabtree, B. F. (2006). The qualitative research interview. *Medical Education, 40*(4), pp. 314-321. doi:10.1111/j.1365-2929.2006.02418.x

Doberstein, D., Agreiter, N., Bäumer, M., Cui, M., Abdollahzadegan, S., Heidari, D., . . . Hoppe, H. U. (2016). CuCoMaG - Group Reflection Support in Role-Playing Environments. *DeLFI 2016 - Die 14. E-Learning Fachtagung Informatik*, (pp. 327-329).

Earp, J., Ott, M., Popescu, M., Romero, M., & Usart, M. (2014). Supporting human capital development with serious games: An analysis of three experiences. *Computers in Human Behavior*, pp. 715-720. doi:10.1016/j.chb.2013.09.004

Eckensberger, L. (1998). Menschenbilder und Entwicklungskonzepte. In H. Keller, *Lehrbuch Entwicklungspsychologie* (pp. 11-56). Bern: Huber.

Emmerich, K., Neuwald, K., Othlinghaus, J., Ziebarth, S., & Hoppe, H. U. (2012). Training conflict management in a virtual environment. *International Conference on Collaboration and Technology* (pp. 17-32). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-33284-5_2

Erdmüller, A., & Jiranek, H. (2010). *Konfliktmanagement. Konflikten vorbeugen, sie erkennen und lösen.* Freiburg: Haufe-Lexware GmbH & Co. KG.

Flynn, R., McKinnon, L., Bacon, E., & Webb, J. (2011). Maritime City: Using games technology to train social workers - some initial results. *International Conference on Entertainment Computing* (pp. 415-418). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-24500-8_55

Fornell, C., & Wernerfelt, B. (1988). A Model for Customer Complaint Management. *Marketing Science, 7(3)*, pp. 287-298.

Gee, J. P. (2007). *Good video games and good learning. Collected essays on video games, learning, and literacy.* New York: Peter Lang.

Gelernter, D. (1985). Generative communication in Linda. *Transactions on Programming Languages and Systems (TOPLAS), 7*(1), pp. 80-112.

Glock, F., Junker, A., Kraus, M., Lehrian, C., Schäfer, A., Hoffmann, S., & Spierling, U. (2011). "Office Brawl" - A Coversational Storytelling Game and its Creation Process. *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology (ACE 2011)* (p. 88). New York: ACM. doi:10.1145/2071423.2071529

Gökel, Ö., & Dağlı, G. (2017). Effects of Social Skill Training Program on Social Skills of Young People. *EURASIA Journal of Mathematics, Science and Technology Education, 13*(11), pp. 7365-7373.

Goodwin, C., & Ross, I. (1990). Customer Evaluations of Responses to Complaints: What's Fair and Why. *Journal of Consumer Marketing, 7*(2), pp. 39-47.

Greif, S. (2008). *Coaching und ergebnisorientierte Selbstreflexion: Theorie, Forschung und Praxis des Einzel-und Gruppencoachings.* Göttingen: Hogrefe Verlag.

Handelsverband Deutschland – HDE e.V. (2020). *Online Monitor 2020.* Retrieved from https://einzelhandel.de/online-monitor

Harbarth, L., Delsing, S., Richtscheid, F., Yücepur, V., Feldmann, F., Akhavanfar, M., Manske, S., Othlinghaus, J. & Hoppe, H. U. (2018). Learning by Tagging - Supporting Constructive Learning in Video-Based Environments. *DeLFI 2018 - Die 16. E-Learning Fachtagung Informatik* (pp. 105-116). Bonn: Gesellschaft für Informatik e.V.

Healey, M., & Jenkins, A. (2000). Kolb's Experiential Learning Theory and Its Application in Geography in Higher Education. *Journal of Geography, 99*(5), pp. 185-195. doi:10.1080/00221340008978967

Heung, V., & Lam, T. (2003). Customer complaint behaviour towards hotel restaurant services. *International Journal of Contemporary Hospitality Management, 15(5)*, 283-289.

Holtgraves, T. M., Ross, S. J., Weywadt, C. R., & Lin, H. T. (2007). Perceiving artificial social agents. *Computers in human behavior, 23(5)*, pp. 2163-2174. doi:10.1016/ j.chb.2006.02.017

Hoque, M., Courgeon, M., Martin, J. C., Mutlu, B., & Picard, R. W. (2013). MACH: My Automated Conversation coacH. *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, (pp. 697-706). doi:10.1145/ 2493432.2493502

Hristidis, V. (2018). Chatbot Technologies and Challenges. *First International Conference on Artificial Intelligence for Industries (AI4I)*, (pp. 126-126). doi:10.1109/ AI4I.2018.8665692

Hubal, R. C., Frank, G. A., & Guinn, C. I. (2003). Lessons learned in modeling schizophrenic and depressed responsive virtual humans for training. *Proceedings of IUI*, (pp. 85-92). doi:10.1145/604045.604062

Hussain, S., Sianaki, O. A., & Ababneh, N. (2019). A survey on conversational agents/chatbots classification and design techniques. In L. Barolli, M. Takizawa, F. Xhafa, & T. Enokido (Ed.), *Web, Artificial Intelligence and Network Applications WAINA 2019. Advances in Intelligent Systems and Computing. 927*, pp. 946-956. Cham: Springer. doi:10.1007/978-3-030-15035-8_93

Ijsselsteijn, W. A., De Kort, Y., & Poels, K. (2013). *The game experience questionnaire.* Technische Universität Eindhoven.

Iten, N., & Petko, D. (2016). Learning with serious games: Is fun playing the game apredictor of learning success? *British Journal of Educational Technology, 47*(1), pp. 151-163. doi:10.1111/bjet.12226

Jain, M., Kumar, P., Kota, R., & Patel, S. (2018). Evaluating and Informing the Design of Chatbots. *Proceedings of the 2018 Designing Interactive Systems Conference*, (pp. 895-906).

Johnson, C. I., Bailey, S., & Van Buskirk, W. L. (2017). Designing effective feedback messages in serious games and simulations: a research review. In *Instructional Techniques to Facilitate Learning and Motivation of Serious Games* (pp. 119-140). Cham: Springer. doi:10.1007/978-3-319-39298-1_7

Johnson, D., Gardner, J., & Perry, R. (2018). Validation of two game experience scales: the Player Experience of NeedSatisfaction (PENS) and Game Experience Questionnaire (GEQ). *International Journal of Human-Computer Studies, 118*, pp. 38-46. doi:10.1016/j.ijhcs.2018.05.003

Jonassen, D., Mayes, T., & McAleese, R. (1993). A manifesto for a constructivist approach to uses of technology in higher education. In T. M. Duffy, J. Lowyck, & D. H. Jonassen, *Designing Environments for Constructive Learning* (pp. 231-247). Berlin, Heidelberg: Springer.

Kapočiūtė-Dzikienė, J. (2020). A Domain-Specific Generative Chatbot Trained from Little Data. *Applied Sciences, 10*(7), p. 2221.

Kerly, A., Hall, P., & Bull, S. (2006). Bringing chatbots into education: Towards natural language negotiation of open learner models. *International Conference on Innovative Techniques and Applications of Artificial Intelligence.* London: Springer. doi:10.1007/978-1-84628-666-7_14

Kickmeier-Rust, M. D., & Albert, D. (2010). Micro-adaptivity: Protecting immersion in didactically adaptive digital educational games. *Journal of Computer Assisted Learning, 26*(2), pp. 95-105. doi:10.1111/j.1365-2729.2009.00332.x

Kiili, K. (2005). Digital game-based learning: Toward an experiential gaming model. *Internet and Higher Education, 8*(1), pp. 13-24. doi:10.1016/j.iheduc.2004.12.001

Kim, P., Hong, J.-S., Bonk, C., & Lim, G. (2009). Effects on group reflection variations in project-based learning integrated in a Web 2.0 learning space. *Interactive Learning Environments, 19*(4), pp. 333-349. doi:10.1080/10494820903210782

Knode, S., & Knode, J.-D. (2011). Using a simulation program to teach leadership. *Proceedings of the 2011 ASCUE Summer Conference*, (pp. 86-92).

Kolb, A. (1984). *Experiential Learning: Experience as the Source of Learning and Development.* Englewood Cliffs: Prentice-Hall.

Krämer, N. C., Bente, G., Eschenburg, F., & Troitzsch, H. (2009). Embodied conversational agents: research prospects for social psychology and an exemplary study. *Social Psychology, 40*(1), pp. 26-36. doi:10.1027/1864-9335.40.1.26

Lamb, R. L., Annetta, L., Firestone, J., & Etopio, E. (2018). A meta-analysis with examination of moderators of student cognition, affect, and learning outcomes while using serious

educational games, serious games, and simulations. *Computers in Human Behavior, 80*, pp. 158-167. doi:10.1016/j.chb.2017.10.040

Lameras, P., Arnab, S., Dunwell, I., Steward, C., Clarke, S., & Petridis, P. (2017). Essential features of serious games design in higher education: Linking learning attributes to game mechanics. *British Journal of Educational Technology, 48*(4), pp. 972-994. doi:10.1111/bjet.12467

Lane, H. C., & Hays, M. J. (2008). Getting down to business: Teaching cross-cultural social interaction skills in a serious game. *Workshop on Culturally Aware Tutoring Systems (CATS)*, (pp. 35-46).

Laurillard, D. (2013). *Rethinking university teaching: A conversational framework for the effective use of learning technologies.* London & New York: Routledge.

Lim, M. Y., Aylett, R., Enz, S., Kriegel, M., Vannini, N., Hall, L., & Jones, S. (2009). Towards Intelligent Computer Assisted Educational Role-Play. *Edutainment 2009*, (pp. 208-219). doi:10.1007/978-3-642-03364-3_27

Linssen, J., de Groot, T., Theune, M., & Bruijnes, M. (2014). Beyond simulations: serious games for training interpersonal skills in law enforcement. *Proceedings of European Social Simulation Association*, (pp. 604-607). Retrieved from http://ddd.uab.cat/record/125597

Malzahn, N., Buhmes, H., Ziebarth, S., & Hoppe, H. U. (2010). Supporting reflection in an immersive 3D learning environment based on role-play. *European Conference on Technology Enhanced Learning* (pp. 542-547). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-16020-2_51

Malzahn, N., Weinbrenner, S., Hüsken, P., Ziegler, J., & Hoppe, H. U. (2007). Collaborative Ontology Development—Distributed Architecture and Visualization. *Proceedings of the German E-Science Conference.* Baden-Baden: Max Planck Digital Library.

Marr, A. C. (2010). *Serious Games für die Informations- und Wissensvermittlung. Bibliotheken auf neuen Wegen.* Dinges & Frick.

Marsh, T. (2011). Serious games continuum: Between games for purpose and experiential environments for purpose. *Entertainment Computing, 2*(2), pp. 61-68. doi:10.1155/2014/358152

Martens, A., Diener, H., Malo, S., Pan, Z., Cheok, A., Müller, W., & El Rhalibi, A. (2008). Games-Based Learning with Computers - Learning, Simulations, and Games. *Transactions on Edutainment I*, pp. 172-190. doi:10.1007/978-3-540-69744-2_15

Mateas, M., & Stern, A. (2003). Façade: An Experiment in Building a Fully-Realized Interactive Drama. *Proceedings of the Game Developers Conference (GDC '03).* San Jose.

Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., van Ruijven, T., . . . Wenzler, I. (2014). The research and evaluation of serious games: Toward a comprehensive

methodology. *British Journal of Educational Technology, 45*(3), pp. 502-527. doi:10.1111/bjet.12067

McCroskey, J. C., & McCain, T. A. (1974). The measurement of interpersonal attraction. *Speech Monographs, 41*(3), pp. 261-266. doi:10.1080/03637757409375845

Menzel, W. (2003). Sprachverarbeitung - ein Überblick. In G. Görz, *Handbuch der künstlichen Intelligenz* (pp. 643-663). München: Oldenbourg Wissenschaftsverlag.

Michael, D. R., & Chen, S. L. (2006). *Serious Games: Games that Educate, Train, and Inform.* Muska & Lipman/Premier-Trade.

Mitchell, A., & Savill-Smith, C. (2005). *The use of computer and video games for learning. A review of the literature.* London: LSDA.

Mock, A. (2017). Open (ed) Classroom–Who cares? *MedienPädagogik: Zeitschrift für Theorie, 28*, pp. 57–65. doi:10.21240/mpaed/28/2017.02.26.X

Mori, D., Berta, R., de Gloria, A., Fiore, V., & Magnani, L. (2013). An easy to author dialogue management system for serious games. *Journal on Computing and Cultural Heritage, 6*(2), pp. 1-15. doi:10.1145/2460376.2460381

Murray, J. H. (2017). *Hamlet on the Holodeck: The Future of Narrative in Cyberspace.* Cambridge, London: MIT Press.

Nacke, L. (2009). *Affective ludology: Scientific measurement of user experience in interactive entertainment.* Karlskrona: Blekinge Institute of Technology.

Naido, S., Ip, A., & Linser, R. (2000). Dynamic goal-based role-play simulation on the web: A case study. *Journal of Educational Technology & Society, 3(3)*, pp. 190-202. Retrieved from www.jstor.org/stable/jeductechsoci.3.3.190

Othlinghaus, J., & Hoppe, H. U. (2016). Supporting Group Reflection in a Virtual Role-Playing Environment. *International Conference on Intelligent Technologies for Interactive Entertainment* (pp. 292-298). Cham: Springer. doi:10.1007/978-3-319-49616-0_30

Othlinghaus-Wulhorst, J., & Hoppe, H. U. (2020). A Technical and Conceptual Framework for Serious Role-Playing Games in the Area of Social Skill Training. *Frontiers in Computer Science, 2*(28). doi:10.3389/fcomp.2020.00028

Othlinghaus-Wulhorst, J., Jedich, A., Hoppe, H. U., & Harrer, A. (2018). Using eye-tracking to analyze collaboration in a virtual role play environment. *International Conference on Collaboration and Technology* (pp. 185-197). Cham: Springer.

Othlinghaus-Wulhorst, J., Mainz, A., & Hoppe, H. U. (2019). Training Customer Complaint Management in a Virtual Role-Playing Game: A User Study. *European Conference on Technology Enhanced Learning* (pp. 436-449). Cham: Springer. doi:10.1007/978-3-030-29736-7_33

Otto, K., Sobiraj, S., Schladitz, S., Vásquez, M. E., Roe, R., & Mabunda, M. B. (2019). Do Social Skills Shape Career Success in the Psychology Profession? *Zeitschrift Für Arbeits-Und Organisationspsychologie A&O, 63*, pp. 88-99. doi:10.1026/0932-4089/a000293

Pivec, P. (2009). Game-based learning or game-based teaching? *British Educational Communications and Technology Agency (BECTA), Report No. 1509*.

Prensky, M. (2000). *Digital game-based learning.* New York: McGraw-Hill. doi:10.1145/950566.950596

Rahim, A., & Bonoma, T. V. (1979). Managing organizational conflict: A model for diagnosis and intervention. *Psychological Reports, 44(3)*, pp. 1323-1244. doi:10.2466/pr0.1979.44.3c.1323

Riggio, R. E. (1986). Assessment of basic social skills. *Journal of Personality and social Psychology, 51*(3), pp. 649-660.

Ritterfeld, U., Cody, M., & Vorderer, P. (2009). *Serious Games: Mechanics and Effects.* New York, London: Routledge.

Romero, M., Usart, M., Ott, M., Earp, J., & de Freitas, S. (2012). Learning through playing for or against each other? Promoting collaborative learning in digital game based learning. *Proceedings of the 20th European Conference on Information Systems (ECIS) 2012.* Barcelona. Retrieved from http://aisel.aisnet.org/ecis2012/93

Salen, L., & Zimmerman, E. (2004). *Rules of Play: Game Design Fundamentals.* Cambridge and London: The MIT Press.

Sandelowski, M. (2000). Combining Qualitative Sampling, Data Collection, and Analysis Techniques in Mixed-Method-Studies. *Research in Nursing & Health, 23*(3), pp. 246-255. doi:10.1002/1098-240X

Schön, D. A. (1983). *The Reflective Practitioner. How Professionals Think in Action.* New York: Basic Books.

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence, 4*(6), pp. 103-108.

Schussler, D., Frank, J., Lee, T.-K., & Mahfouz, J. (2017). Using Virtual Role-Play to Enhance Teacher Candidates' Skills in Responding to Bullying. *Journal of Technology and Teacher Education, 25*(1), pp. 91-120.

Shawar, B. A., & Atwell, E. (2007). Chatbots: Are they really useful? *Ldv forum*, pp. 29-49.

Shawar, B. A., & Atwell, E. S. (2005). Using corpora in machine-learning chatbot systems. *International journal of corpus linguistics, 10*(4), pp. 489-516.

Shen, C., Wang, H., & Ritterfeld, U. (2009). Serious Games and Seriously Fun Games: Can They Be One and the Same? In U. Ritterfeld, M. Cody, & P. Vorderer, *Serious Games: Mechanisms and Effects* (pp. 48-62). New York, London: Routledge.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, pp. 153-189. doi:10.3102/0034654307313795

Smith, C. (2001). Using Social Stories to Enhance Behaviour in Children with Autistic Spectrum Difficulties. *Educational Psychology in Practice, 17*(4), pp. 337-345.

Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Computers & Education, 151*. doi:10.1016/ j.compedu.2020.103862

Solomon, M., Goodlin-Jones, B. L., & Anders, T. F. (2004). A social adjustment enhancement intervention for high functioning autism, Asperger's syndrome, and pervasive developmental disorder NOS. *Journal of autism and developmental disorders, 34*(6), pp. 649-668.

Spafford, S., & Haarhoff, B. (2015). What are the conditions needed to facilitate online self-reflection for cognitive behaviour therapy trainees? *Australian Psychologist, 50*(3), pp. 232-240. doi:10.1111/ap.12076

Spierling, U. (2008). 'Killer Phrases': Design Steps for a Game with Digital Role-Playing Agents. In Z. Pan, A. D. Cheok, W. Müller, & A. El Ehalibi, *Transactions on Edutainment I* (pp. 150-161). Berlin, Heidelberg: Springer. doi:10.1007/978-3-540-69744-2_13

Squire, K., & Jenkins, H. (2003). Harnessing the power of games in education. *Insight, 3(1)*, pp. 5-33.

Stauss, B., & Seidel, W. (2010). Complaint Management. In G. Salvendy, & W. Karwowski, *Introduction to Service Engineering* (pp. 414-432). New Jersey: John Wiley & Sons. doi:10.1002/9780470569627.ch20

Stauss, B., & Seidel, W. (2014). *Beschwerdemanagement: Unzufriedene Kunden als profitable Zielgruppe.* Munich: Carl Hanser Verlag GmbH & Co. KG.

Storp, M. (2002). Chatbots. Möglichkeiten und Grenzen der maschinellen Verarbeitung natürlicher Sprache. *Networx, 25*. doi:10.15488/2916

Susi, T., Johannesson, M., & Backlund, P. (2007). *Serious Games: An Overview. Technical Report HS-IKI-TR-07-001.* Skövde: School of Humanities and Informatics, University of Skövde.

Totty, M. (2005). Better training through gaming. *Wall Street Journal - Eastern Edition, 245(80)*, p. R6.

Turani, A., & Calvo, R. (2006). Sharing Synchronous Collaborative Learning Structures using IMS Learning Design. *7th International Conference on Information Technology Based Higher Education and Training*, (pp. 25-34). doi:10.1109/ITHET.2006.339740

Vaassen, F., & Wauters, J. (2012). deLearyous: Training interpersonal communication skills using unconstrained text input. *Proceedings of ECGBL*, (pp. 505-513).

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), pp. 197-221.

Verpoorten, D., Westera, W., & Specht, M. (2011). Reflection amplifiers in online courses: a classification framework. *Journal of Interactive Learning Research, 22*(2), pp. 167-190.

Wallace, R. S. (2004). *The elements if AIML Style.* ALICE AI Foundation.

Wallace, R. S. (2009). The Anatomy of A.L.I.C.E. In R. Epstein, G. Roberts, & G. Beber, *Parsing the Turing Test* (pp. 181-210). Dordrecht: Springer. doi:10.1007/978-1-4020-6710-5_13

Wang, H., Shen, C., & Ritterfeld, U. (2009). Enjoyment of Digital Games: What Makes Them "Seriously" Fun? In U. Ritterfeld, M. Cody, & P. Vorderer, *Serious game: mechanisms and effects* (pp. 25-47). New York, London: Routledge.

Weinbrenner, S. (2012). *SQLSpaces - A Platform for Flexible Language-Heterogeneous Multi-Agent Systems [Doctoral dissertation].* Munich: Verlag Dr. Hut.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), pp. 36-45.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation.* New York: Freeman.

Wendel, V., Gutjahr, M., Göbel, S., & Steinmetz, R. (2013). Designing collaborative multiplayer serious games. *Education and Information Technologies, 18(2)*, pp. 287-308. doi:10.1007/s10639-012-9244-6

Whitton, N., & Hollins, P. (2008). Collaborative virtual gaming worlds in higher education. *ALT-J, 16(3)*, pp. 221-229. doi:10.1080/09687760802526756

Wiemeyer, J., Kickmeier-Rust, M., & Steiner, C. M. (2016). Performance Assessment in Serious Games. In R. Dörner, S. Göbel, W. Effelsberg, & J. Wiemeyer, *Serious Games* (pp. 273-302). Cham: Springer. doi:10.1007/978-3-319-40612-1_10

Winkler, R., & Söllner, W. (2018). Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis. *Academy of Management Annual Meeting (AOM).* Chicago: USA.

Wisdom, J., & Creswell, J. W. (2013). *Mixed Methods: Integrating Quantitative and Qualitative Data Collection and Analysis While Studying Patient-Centered Medical Home Models.* Rockvill: Agency for Healthcare Research and Quality.

Witmer, B., & Singer, M. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments, 7*(3), pp. 225-240. doi:10.1162/105474698565686

Yee, N. (2006). The demographics, motivations, and derived experiences of users of massively multi-user online graphical environments. *Presence: Teleoperators and Virtual Environments, 15*, pp. 309-329. doi:10.1162/pres.15.3.309

References

Yusoff, A., Crowder, R., Gilbert, L., & Wills, G. (2009). A conceptual framework for serious games. *2009 Ninth IEEE International Conference on Advanced Learning Technologies* (pp. 21-23). Riga: IEEE. doi:10.1109/ICALT.2009.19

Zhang, J., Oh, Y. L., Lange, P., Yu, Z., & Fukuoka, Y. (2020). Artificial Intelligence Chatbot Behavior Change Model for Designing Artificial Intelligence Chatbots to Promote Physical Activity and a Healthy Diet: Viewpoint. *Journal of medical Internet research, 22*(9). doi:10.2196/22845

Zhonggen, Y. (2019). A Meta-Analysis of Use of Serious Games in Education over a Decade. *International Journal of Computer Games Technology, 2019*. doi:10.1155/ 2019/4797032

Zyda, M. (2005). From visual simulation to virtual reality to games. *Computer, 38*(9), pp. 25-32.

DuEPublico

Duisburg-Essen Publications online

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

ub | universitäts bibliothek

Diese Dissertation wird via DuEPublico, dem Dokumenten- und Publikationsserver der Universität Duisburg-Essen, zur Verfügung gestellt und liegt auch als Print-Version vor.

DOI: 10.17185/duepublico/77876
URN: urn:nbn:de:hbz:465-20230512-112544-4

Alle Rechte vorbehalten.